

DOCUMENT RESUME

ED 193 248

TM 800 562

AUTHOR Ward, Annie W., Ed.; And Others
 TITLE Guide for School Testing Programs.
 INSTITUTION National Council on Measurement in Education,
 Washington, D.C.
 PUB DATE [74]
 NOTE 93p.
 AVAILABLE FROM National Council on Measurement in Education, 1230
 17th St., N.W., Washington, DC 20036 (\$4.75).
 EDRS PRICE MF01 Plus Postage. PC Not Available from EDRS.
 DESCRIPTORS Elementary Secondary Education; *Program Development;
 Program Evaluation; *School Districts; Scoring;
 Standardized Tests; Teacher Evaluation; Teacher Made
 Tests; *Testing; Testing Problems; *Testing Programs;
 Test Interpretation; Test Results
 IDENTIFIERS *Test Directors

ABSTRACT

Basically, test directors have four systemwide and building level responsibilities: (1) planning, administration and evaluation of standardized testing programs; (2) application of measurement and evaluation to school problems; (3) classroom testing and evaluation; and (4) external testing programs. In a series of brief memos, measurement and evaluation professionals give advice to the test director on: professional training; advisory committees; test planning; administration; scoring; and applications of test results to guidance and instruction, program evaluation, and teacher evaluation. Other chapters discuss the director's role in training teachers to develop their own tests and administering tests sponsored externally. A final chapter describes and evaluates common methods for reporting individual test results (percentile rank, expectancy tables); and group results (national norms, grade equivalents, criterion referenced tests, matrix sampling). Contributors are: Annie W. Ward; Joseph L. Mazur; Joan K. Bollenbacher; Albert H. Rouse; Gerald S. Hanna; Richard L. Poole; Jack Menne; Margaret E. Backman; and James H. Lawson. (CP)

 * Reproductions supplied by EDRS are the best that can be made *
 * from the original document. *

ED193248

U.S. DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY.

GUIDE FOR SCHOOL TESTING PROGRAMS

Edited by:

Annie W. Ward
Margaret E. Backman
Bruce W. Hall
Joseph L. Mazur

Published by:

National Council on Measurement in Education
206 S. Kedzie, Michigan State University
East Lansing, Michigan 48824

All rights reserved

"PERMISSION TO REPRODUCE THIS
MATERIAL IN MICROFICHE ONLY
HAS BEEN GRANTED BY

Wm Russell
NCME

TO THE EDUCATIONAL RESOURCES
INFORMATION CENTER (ERIC)."

TM800562

TABLE OF CONTENTS

Introduction: The Test Director and Testing Programs in Today's School	1
Chapter I Organization of Testing Services	5
Memo 1 Getting Started	9
Memo 2 Selecting and Using an Advisory Committee	18
Chapter II The Standardized Testing Program	20
Memo 3 Planning the Testing Program	21
Memo 4 Supervising Test Administration	25
Memo 5 Operating a Scoring Service	32
Chapter III Using Results of Standardized Tests	41
Memo 6 Reporting Test Results for Guidance and Instructional Purposes	45
Memo 7 Use of Tests in Program Evaluation	55
Memo 8 The Use of Standardized Tests to Evaluate Teaching	63
Chapter IV Teacher-Made Tests: The Test Director's Role	66
Chapter V Coordinating External Testing Programs	71
Chapter VI Reporting Test Results to the Public	82

PREFACE

This publication has developed in several stages. Initially, the Board of Directors of the National Council on Measurement in Education (NCME) identified the need to prepare some materials for the guidance and support of newly appointed Test Directors of school systems. The need for such materials was discussed at an NCME Board of Directors meeting in the fall of 1970. Warren Willingham of the College Board, who was then editor of Measurement in Education, assumed the responsibility for collecting ideas and organizing them into a tentative outline. In the spring of 1971, the Special Publications Committee was appointed with the following members: Annie W. Ward (Chairman), Bruce Hall and Joseph Mazur, all of the University of South Florida, and Anne Elvery of the Hillsborough Co. (Florida) Schools. The first assignment for this committee was to get the TEST DIRECTORS MEMO off the ground. With the help of Margaret Fleming, Cleveland Public Schools, the tentative outline was expanded and circulated to working test directors for reactions and comments.

As the Special Publications Committee further considered its charge, it was decided to expand the intended audience somewhat to include experienced test directors, as well as those who had been newly appointed. It was also decided to solicit materials from practicing test directors, rather than having a single author prepare all the materials. A further decision was made to publish the materials as they became available, rather than waiting until everything was written and preparing a single publication.

A regular NCME publication, Measurement News, was selected as the outlet for a series of Test Directors Memos, and publication began in the Spring issue of 1973. Ten of these memos had been published through the Spring, 1974 issue. At that point, the idea of collecting the memos into a single booklet was revived. The possibility was also explored of further expanding the audience to include school personnel whose primary job is not that of test director but who are significantly involved in the planning, administration and use of results of the testing program. This monograph is the result.

Many of the original memos are included essentially as they were written. Others were modified or re-written in order to make them applicable to the broader audience, and a few new memos were written to provide for the coverage which the Committee felt was desirable. We hope that the resulting publication can further the development, operation and improvement of testing programs. We feel that the information in this monograph can be useful to a system-wide test director in organizing and operating a system-wide testing program. It should also be useful to superintendents, supervisors of instruction, principals, guidance directors, and teachers in helping them understand what the test director is (or should be) doing and how the testing program relates to the rest of the school program.

Many people have contributed to this endeavor. The individual authors are indicated for each topic. At this point we shall simply list the individuals:

Dr. Anthony J. Allen, University of Rhode Island
Dr. Margaret Backman, ICD Research and Rehabilitation Center, N.Y.
Ms. Joan Bollenbacher, Cincinnati Public Schools
Dr. Gerald Hanna, Kansas State College
Dr. James Lawson, University of Colorado
Dr. Joseph L. Mazur, University of South Florida,
(formerly Cleveland Public Schools)
Dr. John Menne, Iowa State University
Dr. Richard Poole, The National College of Education
Mr. Albert Rouse, Cincinnati Public Schools
Dr. Annie W. Ward, University of South Florida
(formerly Volusia County, Florida, Public Schools)

Special Publications Committee:

Margaret Backman, Bruce W. Hall, Joseph L. Mazur
Annie W. Ward, Chairman

INTRODUCTION

The Test Director and Testing Programs in Today's Schools

Annie W. Ward¹

Today the Director of Testing finds himself in an unenviable position, pulled in diametrically opposed directions by almost irresistible forces. Some of the pressures on the test director are pressures on the entire school system, while others focus specifically on the measurement and testing aspects of the educational endeavor.

Irreconcilable, conflicting demands are made on the testing director continually. He* is frequently put in the position of having to carry out and sometimes defend activities to which he is opposed professionally.

In the United States it has become increasingly popular to charge schools with the responsibility to correct the ills of society. The schools are often seen as the most accessible agent to be held accountable for the correction of society's problems. With the possible exception of reading instruction, testing stands out as the most frequent recipient of criticism both from within and without the educational community. On one hand, testing is viewed as a means to obtain "hard data" showing whether the school system is effective in attaining its ill-defined but multitudinous objectives. On the other hand, the testing activity itself is criticized as contributing to, if not creating, many of society's ills, such as the stratification of social status, limitation of opportunities for upward mobility, and invasion of privacy. (Gross, 1962, Hoffman, 1962, 1967; Holtzmann, 1971; Kohl, 1967).

In its short history, the educational testing movement has undergone several periods of criticism and examination. Early criticism was directed primarily against overenthusiastic, uncritical interpretation and use of test results. (e.g., NEA Joint Committee on Testing, 1962). These criticisms led to increased awareness of the need to train test users and to encourage test authors and publishers to provide more information for the test user. (American Psychological Association *et. al.*, 1966).

¹Many of these ideas were contributed by Dr. Margaret Fleming, Cleveland Public Schools.

*In this enlightened decade some of the authors considered referring to the director as "she" throughout the publications. while others gave some thought to "he/she." However, because of the reverse sexism of the first and the awkwardness of the second, we bow to tradition and continue with the inclusive "he."

Other more recent criticism has had a different focus. Part of it has been a reaction to the "dehumanizing" nature of quantification; part is a philosophical denial of the existence of individual differences, with an implied renewal of the nature-nurture controversy. (Black, 1963; Blake, 1970; Hunt, 1961). Let us hope that the ensuing discussions and debates in the educational community will take us another step forward in developing and using tests more effectively and more wisely.

Measurement experts have long been aware of the shortcoming of available tests as a means of observing and measuring behavior. (e.g., Womer, 1967). Unfortunately, the precautionary notes which experts have sounded have been amplified by the critics in such a way that the general citizenry could very well conclude that testing is of little value and, perhaps, even dangerous. It is difficult to counsel for the proper degree of caution in the use of test data without running the risk of total rejection of all tests.

Testing is the most efficient method of securing information that is valid, reliable, and relevant for a broad spectrum of objectives. It serves as a monitoring device that provides information about educational needs and products. It provides a means to describe student behavior in terms that are readily communicable. When used appropriately, testing can provide an information base for educational decision-making related to students, programs, instruction, and curricula. It is basic to the conduct of any meaningful evaluation in which discrepancies between anticipated and actual attainment of objectives can be determined. Along with systematic observation and historical records, testing provides one of the most defensible starting points for a purposeful improvement of education.

In recent years, there has been considerable emphasis on defining the objectives of instruction. Ralph Tyler (w. Smith, 1942, 1965) has long emphasized the need for specifying the goals of education in terms that would be useful in the classroom. Bloom *et. al* (1956) helped to clarify and categorize measurable objectives of instruction. This background was essential to the use of appropriate procedures to assess what happens to the learner as a result of his experience in school. (see also Popham, 1971 and Bloom *et. al.*, 1971).

Many other advances have been made in improving educational evaluation. (Chauncey, 1965). Technical competencies have improved. Fast, efficient scoring machines have taken much of the drudgery out of test processing. Computers provide a wide variety of analyses at command. The use of sophisticated statistical techniques has enabled us to understand better what a given test is measuring. Many more people have now been trained to handle the intricacies of measurement. In addition to paper and pencil tests, we have developed a wide variety of assessment techniques, such as situational tests, observation schedules, and rating scales for behavior and pupil products.

The cornerstone of the assessment program in most school systems is the standardized testing program. Its basic function is to provide information in an objective, systematic fashion, data from which comparisons can be made to other relevant groups. Results from standardized tests have been used in the following ways:

1. *As a screening device to identify students with special instructional needs.*
2. *To identify the need for special courses and programs*
3. *To provide data for instructional and institutional decision making.*
4. *To identify areas of the school program which need further study.*
5. *As an aid to personal, educational, and vocational planning.*

Difficulties arise when these applications are broadened to include some which are not defensible; for example, when an attempt is made to use a survey achievement test battery to diagnose specific learning difficulties of an individual student. Or when test scores are used to assign students to a program or "track" with little attention given to the evaluation of the program or the placement of the student. Or when an attempt is made to evaluate the effectiveness of a program or teacher solely on the basis of scores on tests.

It is important that all who use test data understand what the data mean and that they are aware of the limitations of the data and the precautions which must be used. Insuring proper interpretation is a formidable challenge because it involves people both in and outside of education and involves communication to diverse audiences, each with different points of view and levels of knowledge.

Thus, the test director of a school or school system stands at the center of both controversy and challenge. As a member of the staff in a school system, he is the expert-in-residence with regard to testing and measurement efforts of the system. Coordination of the program, selection of tests, and development of various data-gathering instruments represent typical responsibilities. At the lowest level of skill the mechanics of test distribution, administration and scoring dominate his time. At the most professional level, the test director is an important contributor and staff specialist in planning, evaluation, and development of instructional programs. The test director performs very significant services by providing the operational capability to gather, analyze, and interpret needed data. The training and competencies of a test director provide a means to develop a credible information base for planning and development of school programs.

Two fairly recent developments have expanded the responsibilities of many test directors. One of these is the requirement for evaluation information for such Federal programs as ESEA I, III, IV and VII. The other is the awakened interest in information-based decision making which has grown out of the introduction of "management information" and "program, planning, budgeting systems." Too frequently in the past, testing has not been recognized by administrators as a basic ingredient of decision making, but that is changing.

In only a few circles has testing been viewed as an important information base upon which educational planning and development can rely. Frequently, testing has operated in isolation from such other important school functions as evaluation, instructional planning and curricular development.

The importance of the decisions which educators must make necessitates dependable data which can be used as a basis for those decisions. The organization and operation of a regular systematic program of carefully chosen tests is the first step. Experience and training in interpreting the data is another necessary requirement.

CHAPTER I - ORGANIZATION OF TESTING SERVICES

Annie W. Ward and Joseph L. Mazur

An effective testing program requires two levels of organization: (1) System-wide level and (2) Building level. Very large systems may also need an intermediate (district) level. The key person in the organization is the system-wide test director. He is the person that makes the organization work. Furthermore, the organizational location of the system-wide test director can have important implications for the way he functions and for the way the testing program is viewed and used by others in the school system.

The Job of the Test Director

Before considering the matter of organization, it should be helpful to look at the functions which test directors perform. These functions are common to all test directors, whether their assignment is system-wide or limited to a single school. The difference is in the level of responsibility and emphasis.

One way of looking at the role of the test director is presented in Figure 1. A note about the procedure used in developing this chart should help in interpreting it. The first step was to list the various tasks of a test director. Then each of the tasks was categorized as to "area of responsibility" and "function."

The list of tasks suggested four broad areas for which test directors are generally responsible: (1) standardized testing programs, (2) application of measurement and evaluation to various school problems, (3) classroom testing and evaluation, and (4) special testing programs. The processes by which test directors carry out these responsibilities may be grouped into three functions: Planning, Management, and Professional-Technical Consultation. Planning involves assessing needs, formulating specifications and policies, designing the testing programs, and relating the testing programs to other programs of the school system. Management includes the specific operations of acquiring, storing, and distributing materials; scheduling and supervising test administration; preparing data summaries and reports and providing for data storage and inservice training, instrument evaluation and development, interpretation of data and instructional evaluation.

Figure 1 shows the relationship between the major areas of responsibility and the basic functions of the test director. The intersection of the columns and rows identifies specific activities that are associated with each combination of responsibilities and function. Some of the cells are blank because no activities were easily identified at those intersect points. The chart helps to clarify the dimensions of the job of the test director. It also helps to clarify the special competencies associated directly with testing and measurement operations in a school system.

	Standardized Testing Program .100	Application of Measurement in Evaluation .200	Classroom Application of Measurement .300	Special Testing Program .400
Planning	Assessing Needs 1.11	Assessing Needs 1.21		
	Formulating Specifications for Testing Programs 1.12	Formulating Specifications for Instruments 1.22		Developing Policies Regarding Special Testing Programs 1.42
	Formulating the Testing Program 1.13	Formulation of Data Collection Plan 1.23		Selecting Special Programs for Participation 1.43
	Relating the Testing Program to Other Departments 1.14	Integrating Special Assessments with Other Programs 1.24	Integrating Classroom Testing with Other Testing Programs 1.34	Integrating Special Testing Programs with Other Programs 1.44
Managing	Acquiring Materials 2.11	Acquiring Materials 2.21		Collecting and Accounting for Money and Materials 2.41
	Scheduling the Administration 2.12	Scheduling the Data Collection 2.22		Scheduling the Administration 2.42
	Supervising the Administration 2.13	Supervising the Data Collection 2.23		Selecting and Training Administrators 2.43
	Providing Scoring Services 2.14	Providing Scoring Services 2.24	Providing Scoring Services 2.34	Sending Materials to Scoring Center 2.44
	Providing Summaries and Analyses of Data 2.15	Preparing Data Summaries and Analyses 2.25	Reporting 2.35	Receiving and Distributing Reports 2.45
	Storing & Retrieving Data 2.16	Storing Data for Retrieval 2.26	Developing and Maintaining Item Banks 2.36	
Professional - Technical Activities	Consulting 3.11	Consulting 3.21	Consulting 3.31	
	Conducting In-Service Training 3.12	Conducting In-Service Training 3.22	Conducting In-Service Training 3.32	Conducting In-Service Training 3.42
	Evaluating & Developing Instruments 3.13	Preparing Special Instruments 3.23		
	Interpreting Data 3.14	Interpreting Data 3.24		Interpreting Data 3.44
	Evaluating the Program 3.15	Evaluating the Instruments 3.25	Evaluating the Services 3.35	Evaluating the Program 3.45

Figure 1 BASIC FUNCTIONS AND TASKS OF A SCHOOL SYSTEM DIRECTOR OF TESTING

Organizational Location

While Figure 1 is useful in delineating the scope of activities of the job, it is not the whole picture. The organizational and inter-personal context of each situation are critical factors in determining the ways in which the testing service functions. Simply put, the place occupied by the testing service within the organizational structure of the school system will determine to a large degree the effort that is given to selected activities within various cells of Figure 1 and the way the testing service is viewed by others.

Test directors may be located in various departments within the school systems. In most systems they are housed within one of the three divisions: administration, instructional services, or pupil personnel services. There are both advantages and disadvantages in each of these locations. Some of these are discussed below:

Administration. The superintendent and his staff have a continuous need for data as a basis for decision making. If the testing office is a part of a Research Bureau which reports directly to the superintendent, the use of test data in decision-making can be maximized. Such an arrangement may be able to serve all data needs of the entire school system. Unfortunately, the Research Bureau, if present at all, has only a limited role in many school systems. Its activities may emphasize the collection of information related to enrollment, attendance, salary distribution, and the like. If this is the case, the test director's effectiveness will be limited in this organizational location.

Instructional services. Achievement testing must obviously be relevant to the instructional program and test results need to be used in instructional planning. Locating the test director in the instructional division can facilitate the kind of communication necessary to make the testing program relevant for the curriculum and can make the curriculum more responsive to needs identified through testing. However, there is the possibility that the testing program could become so curriculum-bound that the need for broader kinds of information would be ignored. A balance is needed between effective use of test information and needs of specific divisions within a school system.

Pupil personnel services. In some school systems the testing program is a part of pupil personnel services. Frequently, the director of testing is also responsible for pupil personnel services and many kinds of test data are used in counseling and individual decision-making. As a part of the pupil personnel staff, however, the test director may be viewed apart from instructional and administrative information needs. He may find himself in the position of having to persuade instructional and administrative personnel of the need to consider test data as a basis for curriculum and administrative planning.

It is essential that the testing program be viewed as a service for the entire school system, and, therefore, the needs of the entire system must be considered in establishing and operating the program. Each of the possible organizational locations for the testing program will facilitate consideration of some needs, but hamper consideration of others. The test director must capitalize on the advantages while remaining aware of the difficulties which arise from the organizational setting in which he finds himself. If the organizational location makes it impossible to function effectively, he will have to try to change that location.

Relationship to the Total School System

The contribution that the testing service makes to the effective operation of a school or school system is directly affected by how important other departments perceive the testing service to be for their own operations. If objective data are valued as a basis for decision-making, the testing service can respond in creative ways that go beyond the mere processing and reporting of test data through general bulletins. A handtailored approach can be developed for the special needs of various departments. The testing service, in this case, can play an integral part in decision making.

To the extent that he views his role as limited to that of data collector and impartial reporter, the activities of the testing director will not be closely integrated with operations of other departments. On the other hand, if the test director and his staff attempt to function in a more active and supportive manner, it is possible for the testing

service to initiate activities that lead to highly relevant, useful information systems designed to meet the varying needs of the school system.

At the school building level, the test director may be a guidance counselor, the principal, or a classroom teacher; this means that the operation of the testing program is a minor part of the responsibility for that person. Therefore, it is very important that the system-wide director organize testing activities in such a way that the job of the school director is facilitated and that as much of the work is carried out centrally as possible. However, there are some aspects of the testing program which must be handled by someone in the school: coordinating the testing schedule with the rest of the school schedule, training teachers to administer tests or arranging for test administration by counselors or others, distributing and collecting test materials, distributing reports of results and helping to interpret them. The system-wide director can provide help and consultation, but most of these operations must be carried out by someone within the school -- the "building test director."

While the focus of this monograph is on the system-wide test director, the role of the building director is also recognized and specific suggestions for people who serve that function are included in many of the materials.

The System-Wide Test Director

The work of the system-wide test director involves certain technical competencies: Knowledge of basic measurement concepts, test selection skills, test development skills, and test interpretation skills. However, it also involves administrative and interpersonal skills. The system-wide test director must organize a complex operation so that tests are selected, purchased, distributed, administered, scored, and reported on a time schedule that makes it possible for school system decision makers (principals, teachers, counselors, supervisors) to use the results. Interpersonal skills are required because much of the test director's job involves working with and through other professionals.

In order to get an overview of the job of the system-wide test director, let us pretend we have been asked by a newly appointed test director for help in getting organized so that he can become functional as quickly as possible. The rest of this chapter consists of specific memos for that newly appointed test director and deal with the essential activities he must carry out to get started. Experienced test directors may wish to study the memos as a check on their own activities.

MEMO I: GETTING STARTED

Joan Bollenbacher, Albert H. Rouse, Jr.

Cincinnati Public Schools

This memo is intended for the newly appointed test director. In retrospect, it might have been more useful had it been called "Memo to a Candidate for Test Director," but since the series is intended to help rather than to discourage new appointees, perhaps the emphasis is appropriate.

If the title of the new appointee is "test director," that fact alone is unusual. In 69 large cities in the United States, only 11 have a "director" or "supervisor of testing." An additional 14 have "tests" or "testing" in their titles, but it is combined with some other activity such as evaluation or guidance and pupil personnel services. In the other 44 cities the testing function is subsumed under other areas such as "special services," "research and development," "educational investigation and measurement," "planning and evaluation," "assessment and accountability," etc.

Obviously, the job of test director can well be described as a "mixed bag." Any new appointee in a public school setting, therefore, is faced with many questions, and his or her ability to get definitive answers may mean the difference between success or failure on the job. In any event, knowing some of the questions, if not all of the answers, is vital to getting off to a good start. This memo will raise some of those questions and then offer comments and suggestions.

WHAT ARE THE GOALS AND OBJECTIVES OF THE SCHOOL SYSTEM?

It is hoped that the new test director either knows in advance or has asked for detailed answers to this question. He will want to know if the Board of Education and/or the Superintendent have addressed themselves to this issue, not only for testing but indeed for the school system itself. Ideally, the goals and objectives would have been developed in consultation with parents, pupils, teachers, and interested persons in the community and are in written form. Rarely, however, is this ideal attained. Paradoxically, the less systematic the approach to this problem, the more important it is for the new test director to know where testing fits into the total picture.

WHAT ARE THE PURPOSES OF TESTING?

However explicit are the purposes of testing in the mind of the new director, he would be well advised to learn as quickly as possible whether his purposes match those of the Superintendent and Board. If there is an organization chart, a glance at it will tell much. Whether the director's job has a line or staff relationship to the administrative officers is a consideration but probably not as important as where testing is located in the total organization. As noted above, rarely is testing given department status; most commonly it is a unit of another department or division.

Testing, of course, serves many purposes, but its place in the organization and its primary purpose probably are closely related. These purposes as they are related to the organization chart might be described as follows:

1. *To provide data to aid in administrative and instructional decisions for the school system (probably combined with research).*
2. *To provide data for counseling students (probably combined with counseling and guidance services).*
3. *To provide data to aid teachers in the instruction of pupils (probably combined with evaluation or placed in the department of instruction).*
4. *To provide data for reporting to the public (probably a part of an accountability unit).*

Depending upon his orientation, the new test director could find himself in an uncomfortable position. For example, if he is guidance oriented and the testing division is in an accountability unit, there indeed could be a "conflict of interest." Of course, the same data can be used in counseling of students and for accountability purposes, but it is obvious that the test director should understand in advance what the primary purpose is; otherwise he could find that his ideas of the purpose of testing are incompatible with those of the administration.

WHY HAS A NEW TEST DIRECTOR BEEN APPOINTED?

As the new appointee is struggling with goals, objectives, and purposes, he also should be asking why a new test director is being appointed. It is likely that one of these three circumstances will prevail:

(1) There has been a long-established tradition of testing in the school system with a well-known, well-liked director who is retiring. In this situation the new appointee will likely be compared frequently with his predecessor and be reminded from time to time that "we didn't do it that way last year." A new test director probably will wish to

Institute some changes, but when he is following a winner, he will wisely move with care.

(2) There has been a tradition of system-wide testing with a director who is leaving abruptly after only a few years in the job. It would be prudent of the new appointee to learn why the former director left. Customarily the reason for a vacancy is made clear in the interview if it is not widely known. If the new director must make discreet inquiry after he has accepted the job, he has a problem which no memo can solve.

(3) There is no system-wide testing program and, thus no previous director. Here the new director has an open field with no tradition and few preconceived ideas about testing. While this situation appears to have many advantages, the new appointee will be faced with a huge task. He will have the problem of orienting the teachers, principals, and community to the testing program and what it means; he will have to establish some plan for inaugurating the testing program, most likely by various committees; and he will have to work out countless administrative details. Future memos in this series will be helpful. In the meantime, the new director can only do his best.

WHAT CRITERIA WERE USED IN SELECTING A TEST DIRECTOR?

Any new test director would be wise to ask himself, "Why me?" Some school systems have established criteria for each position which are published widely when a vacancy occurs. In this instance, the candidate can compare himself with the criteria and very likely come to a valid conclusion as to why he was selected. There are other situations, however, where the primary selection instrument seems to be a crystal ball. It is here where the new director will need to ask the most searching questions of himself. If he is honest in his evaluation, he will have good insight into his acceptance and possible future success in the school system.

WHAT DOES A TEST DIRECTOR NEED TO KNOW?

The fact that he was hired should give the new test director some degree of confidence that he has the background and skills to do the job. But testing is not done in a vacuum. No matter how well-prepared he is, the new test director will be surprised at how frequently his aspect of testing appears to be at the heart of a controversy and how often he is asked to provide information on short notice. The source of the problem may be 2,000 miles away or at the nearest elementary school. It may be an article in the New York Times Magazine severely criticizing tests; or a meeting of the local Chamber of Commerce where the suggestion is made that tests should be used to rate teachers; or a resolution from the local teacher's organization that all testing be abandoned; or a

contingent of concerned parents objecting to a particular test; or an inquiry from a Board of Education member who has heard about "Management by Objectives."

WHERE DOES A TEST DIRECTOR GET INFORMATION?

Not even an experienced test director, let alone a "new" one, can be expected to have ready answers to such widely varied problems as those described above, but he should have immediately available sources for help. Sometimes help may be secured from a professional publication, from a call to a colleague in a professional association (e.g., the National Council on Measurement in Education) or from a test publisher's advisory staff.

In a relatively brief memo such as this, it is not possible to provide a complete list of information sources, but a representative list may be helpful. It should be noted that no exhaustive search was involved in compiling these lists; rather they were gleaned from the contents of the bookcases in the offices of the authors. Consequently, it is likely that some authoritative source is not included, but such an omission should cause no concern when it is remembered that the list is intended to illustrate the scope of the new test director's possible needs.

(1) Professional Associations. In addition to joining the National Council on Measurement in Education, the new test director should consider membership in other professional associations according to his particular needs. Several of these follow:

National Council on Measurement in Education
(NCME)
Office of Evaluation Service
Michigan State University
East Lansing, Michigan 48823

American Educational Research Association
126 Sixteenth Street, N.W.
Washington, D.C. 20036

Association for Measurement and Evaluation Guidance
(AMEG)

American Personnel and Guidance Association
(APGA)
1607 New Hampshire Avenue, N.W.
Washington, D.C. 20009

American Psychological Association
1200 Seventeenth Street, N.W.
Washington, D.C. 20036

(2) Newsletters, Special Publications, and Professional Journals.

Each of the associations listed above sponsors publications which will be useful to the new test director. Among them are MEASUREMENT NEWS (NCME); JOURNAL OF EDUCATIONAL MEASUREMENT (NCME); AMERICAN EDUCATIONAL RESEARCH JOURNAL (AERA); REVIEW OF EDUCATIONAL RESEARCH (AERA); EDUCATIONAL RESEARCHER (AERA); MEASUREMENT AND EVALUATION IN GUIDANCE (AMEG). Another important journal is EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT, Box 6907, College Station, Durham, North Carolina 27703.

(3) Useful References. Below are listed some references which range from the popular and practical to the technical and theoretical. As noted above, they are intended to reflect the spectrum of demands upon the test director.

American Psychological Association. Standards for Educational and Psychological Tests and Manuals. Washington, D.C. 1974.

Anastasi, Anne (Ed.) Testing Problems in Perspective. Washington, D.C.: American Council on Education, 1966.

Bauernfeind, R.H. Building A School Testing Program, Boston: Houghton Mifflin Company, 1969.

Bloom, B.B., Hastings, J.T., and Madaus, G.F. Handbook on Formative and Summative Evaluation of Student Learning. New York: McGraw-Hill, 1971.

Bloom, B.S., and Krathwohl, D.R. Taxonomy of Educational Objectives: Cognitive Domain. New York: David McKay Co., Inc. 1956.

Browder, Leslie H., Jr. Emerging Patterns of Administrative Accountability. Berkeley, Calif.: McCutchan, 1971.

Buros, Oscar Krisen (Ed.) The Seventh Mental Measurements Yearbook. Highland Park, N.J.: The Gryphon Press, 1972.

Cronbach, L.J. Essentials of Psychological Testing, Third Edition. New York: Harper and Row Publishers, 1970.

Ebel, Robert L. Essentials of Educational Measurement. Englewood Cliffs, N.J.: Prentice Hall, Inc. 1972.

Ebel, Robert L.; Noll, Victor H.; and Bauer, Roger M. (Ed's). Encyclopedia of Educational Research, Fourth Edition. London: The MacMillan Company, 1969.

Guilford, J.P. Fundamental Statistics in Psychology and Education. New York: McGraw-Hill, 1965.

- Gulliksen, Harold. Theory of Mental Tests. New York: John Wiley & Sons, Inc., 1955.
- House, Ernest R. School Evaluation: The Politics and Process. Berkeley, Calif.: McCutchan, 1973.
- Lindquist, E.F. Educational Measurement. Washington, D.C., American Council on Education, 1951.
- Lyman, Howard B. Test Scores and What They Mean, Second Edition. Englewood Cliffs, N.J.: Prentice-Hall, Inc., 1971.
- Mehrens, W.A. and Lehmann, I.J. Measurement and Evaluation in Education and Psychology. New York: Holt, Rinehart, and Winston, Inc. 1973.
- Popham, W. James (Ed.) Criterion Referenced Measurement: An Introduction. Englewood Cliffs, N.J.: Educational Technology Publications, 1972.
- Riphey, Robert M. Studies in Transactional Evaluation. Berkeley, Calif.: McCutchan, 1973.
- Russell Sage Foundation. Guidelines for the Collection, Maintenance, and Dissemination of Pupil Records. New York: Russell Sage Foundation, 1970.
- Thorndike, Robert L. (Ed) Educational Measurement, Second Edition. Washington, D.C.: American Council on Education, 1971.
- Thorndike, R.L., and Hagen, E. Measurement and Evaluation in Psychology and Education, Third Edition, New York: John Wiley and Sons, Inc. 1969.
- Whittle, Dean K. (Ed.) Handbook of Measurement and Assessment in Behavioral Sciences. Reading, Mass.: Addison-Wesley Publishing Company, 1968.

Note: The new test director will have his own library of textbooks on educational measurement, statistics, and research. He will find it helpful to refer to at least three or four on a topic when he is faced with the problem of providing a simple explanation of a fairly complex topic to a layman.

(4) Publishers---Materials and Services. Whether the new test director is administering a program of norm-related or criterion-related tests, he will need to know what materials and services are provided by the test publishers. He would do well to have his secretary (let us hope he has one!) check the file to be sure that it contains all the recent test catalogs. A perusal of these can give the new director a quick overview of the various kinds of tests which are available, what types of scoring services are offered, and what kind of professional services and materials are provided. The new test director will find the major publishers to be real professionals in the field and frequently they can provide free counsel and assistance in solving various testing problems. Most of them produce special bulletins and interpretive guides; some offer in-service workshops.

Below are listed the addresses of seven publishers. This is not a complete list but one that should prove helpful. Most of the catalogs list the names of the professional staff of the test division and the local test representatives.

California Test Bureau/McGraw-Hill
Del Monte Research Park
Monterey, California 93940

Educational Testing Service
Princeton, New Jersey 08540

Harcourt Brace Jovanovich, Inc.
757 Third Avenue
New York, New York 10017

Houghton Mifflin Company
110 Tremont Street
Boston, Massachusetts 02107

Personnel Press, Inc.
20 Nassau Street
Princeton, New Jersey 08542

The Psychological Corporation
304 East 45th Street
New York, New York 10017

Science Research Associates, Inc.
259 East Erie Street
Chicago, Illinois 60611

(5) External Testing Programs. Although the test director may have no direct responsibility for testing programs sponsored by outside agencies, parents and teachers probably will ask him for information from time to time. If the office files do not contain up-to-date information on these several agencies and the programs they sponsor, the new test director will want to write to the following addresses:

American College Testing Program
P.O. Box 168
Iowa City, Iowa 52240

Educational Records Bureau
160 Speen Street
Framingham, Mass 01701

College Entrance Examination Board
Box 592
Princeton, New Jersey 08540

General Educational Development Testing Service
American Council on Education
1785 Massachusetts Avenue
Washington, D.C. 20036

National Assessment of Educational Progress
300 Lincoln Tower
1860 Lincoln Street
Denver, Colorado 80203

National Merit Scholarship Corporation
990 Grove Street
Evanston, Illinois 60201

WHAT ARE THE ENCOURAGING WORDS FOR THE NEW TEST DIRECTOR?

The authors of this memo hope that the new director is not overwhelmed by all this detail. If the authors' experiences are indicative, the director will find the job exciting and challenging; he will complain that the days are all too short; and he will be happy that across the country those who have chosen testing for their profession are the kinds of persons he will want as friends. Good luck!

REFERENCES

- American Psychological Association; American Educational Research Association; National Council of Measurement in Education. Standards for educational and psychological tests and manuals. Washington: American Psychological Association, Inc., 1966.
- Blake, Elias Jr., Test information as a reinforcer of negative attitudes towards black Americans. Proceedings of the 1970 invitational test-int conference. Princeton: Educational Testing Service, 1970.
- Black, H., They shall not pass. New York: William Morrow & Company, 1963.
- Bloom, B.S. (ed.) Taxonomy of educational objectives, Handbook 1: cognitive domain. New York: Longmans Green & Co., 1956.
- Bloom, B.S., J.T. Hastings, and G.F. Madaus, Handbook on formative and summative evaluation of student learning. New York: McGraw-Hill, 1971.
- Chauncey, Henry, Progress in testing, 1965. In Gronlund (ed.) Readings in measurement and evaluation. New York: The MacMillan Co., 1968.
- Gross, M.J., The brain-watchers. New York: Random House, 1962.
- Hoffman, Banesh, Tyranny of testing. New York: The Crowell-Collier Press, 1961.
- Hunt, J. McV., Intelligence and experience. New York: Random House, 1962.
- Holtzman, W.H., The changing world of mental measurement and its social significance. American Psychologist, Vol. 26, No.6, June, 1971.
- Joint Committee on Testing, NEA, Testing, testing, testing. Washington: National Education Association, 1962.
- Kohl, Herbert, Thirty-six children. New York: New American Library, 1967.
- Mager, R.F., Preparing instructional objectives. Belmont, Calif. Fearon Publishers, 1962.
- Popham, W.J., An evaluation guidebook: a set of practical guidelines of the educational evaluator. Los Angeles: The Instructional Objectives Exchange, 1971.
- Tyler, R.W., The curriculum in testing problems in perspective. New York: American Council on Education, 1965.
- Tyler, R.W. and Eugene Smith, Appraising and recording student progress Vol. III. New York: Harper & Brothers, 1942.

MEMO NO. 2: SELECTING AND USING AN ADVISORY COMMITTEE

Joan K. Bollenbacher, Cincinnati Public Schools

The selection of a test advisory committee is one of the first steps in establishing a successful testing program. The results of a broadly-based testing program will affect in one way or another students, parents, teachers, counselors, curriculum specialists, administrators, Board of Education members, and many persons in the community; therefore, the need for a widely representative committee is obvious.

Frequently in large school systems there will be a standing system-wide test advisory committee of 10 to 15 members, probably chaired by the director of testing or research. Additional test committees may be organized at the local school level or by areas. Smaller school systems may have only one committee which serves all schools. Whatever the committee organization, some procedures are common to all.

Appointment of the Committee. Although a counselor, a test director, or a teacher may have recognized the need for a test advisory committee, the superintendent or principal may need to be persuaded as to the advisability of appointing such a committee and, in turn, will need to select a person to take the leadership in establishing it. He probably will indicate whether the chairperson is to be a teacher aided by a counselor or test director to give technical assistance, or whether a counselor or test director will chair the committee. Once the committee members have been selected, it would be well for a formal invitation to be sent to each person by the chief administrator.

Charge to the Committee. When the committee meets for the first time, the chief administrator should brief them so that they have a clear understanding of their duties and responsibilities. If a testing program is being reviewed and new tests selected, it is essential that guidelines be established so that the committee knows at the outset who will have the final decision in selecting the tests. This procedure is especially important to avoid possible controversy later if consensus cannot be reached.

The Work of the Committee. If objectives of the testing program are not well established, the first task of the committee will be to develop a statement of objectives for approval by the chief administrator. If the objectives have been established, the committee will need to review them as to their appropriateness and make revisions where necessary.

If a review or revision of a system-wide testing program is involved, the committee will next need to make careful plans for the work ahead. Advice on the technical matters pertaining to the review should be sought from the measurement specialists on the committee or from outside consultants if need be. It is assumed that the specialists will have full

knowledge of the Standards for Educational and Psychological Tests, but they, in turn, should discuss the appropriate sections of the Standards with the committee. Criteria for the review will need to be established and checklists developed for evaluating the various instruments under consideration. In a large program where many test batteries are being reviewed, ad hoc sub-committees can be appointed for preliminary screening to eliminate instruments which clearly fail to meet the criteria. The proceedings for each meeting of the committee(s) should be recorded with careful documentation of the procedures. When the tests or test battery are finally selected, a written report explaining the entire procedure and describing the rationale for the final selection should be presented to the chief administrator for his approval and adoption by the Board of Education where required.

When the Test Advisory Committee operates on a continuing basis, their work does not cease when the tests are adopted and the testing program organized. Rather, the committee will provide continuing evaluation of the program as it relates to test administration; use of results; and interpretation to students, parents, and the community at large. Ultimately, the committee should provide an evaluation of the program's effectiveness in meeting the stated objectives.

CHAPTER II - THE STANDARDIZED TESTING PROGRAM

The major part of a system-wide testing service is the program of standardized testing. The regular, systematic collection of test data which can be referred to national norms is basic to the entire data collection system for a school system. This does not mean that testing services should be limited to standardized tests. Locally developed tests and tests developed to evaluate special projects are also necessary; so are the tests which classroom teachers use in assessing the progress of individual students. However, we should not lose sight of the importance of the basic standardized testing program.

The usefulness of the standardized testing program depends on many factors: (1) the care with which the program is planned, (2) the attention to standardized procedures in administering the tests, (3) the accuracy and efficiency of the scoring and reporting of results, and (4) the way the results are interpreted and used. This chapter includes memos dealing with the first three. The use of test results is so important and so complex that an entire chapter is devoted to it. (Chapter III).

MEMO NO. 3: PLANNING THE TESTING PROGRAM

Albert H. Rouse, Jr.

Cincinnati Public Schools

A school testing program without planning and direction wastes money, or possibly worse, gets out of bounds. First, it becomes stretched beyond its purposes, then it is exploited, and finally scorned. Often the test results are not used when they could have far-reaching positive effects, or they are used in ways that are unjustifiable and never intended. A testing program is continuously in a state of planning and adjustment to meet educational needs. A review of the basics of planning a testing program should be helpful. Those of us with responsibility for testing programs long in existence or one just beginning are going to remain on top of our position only when we won't "let George do it."

WHY A TESTING PROGRAM?

The reason for a testing program comes from expressed needs. Students need to make educational decisions. Teachers, principals, administrators, and parents are likewise involved in educational decision making. Tests are a major tool in helping to supply information for this task.

Students need to obtain information about themselves. Standardized test results can provide comparisons that go beyond the immediate experience of the student. They can help him evaluate aspects of himself with peer groups beyond his school and neighborhood. They can help him develop a self concept that will not as readily become threatened when exposed to a different school, a different neighborhood, and a changing environment.

WHO SHOULD BE INVOLVED?

Testing programs do not function in a vacuum. It would seem most practical in planning a program to involve representatives from the groups who will administer the tests and/or use the results. Including representatives from groups who will be otherwise involved with the testing program serves two purposes. First, they will understand more clearly the purposes and thus help the program gain acceptance. Second, representation from each group increases the spectrum of knowledge and skill in the planning process. Beyond these purposes, a well-chosen planning committee may stay on as a valuable advisory group.

DEFINING OBJECTIVES OF A TESTING PROGRAM

Major consideration should be given to the clear statement of objectives for a testing program. There is danger here of trying to emphasize too many objectives in an attempt to meet the needs of all potential users. It is doubtful if any school system could afford the testing program "dreamed up" without restriction. Objectives must be put in priority. One suggestion to consider is to define the prime user and to tailor the testing program to meet that user's needs. It is quite likely that the results would still be useful to the other users. Over time there may be need to redefine the prime user.

SELECTING APPROPRIATE INSTRUMENTS

Once the objectives of the program are defined and put in priority order, consideration should be given to the selection of the appropriate tests or measurement tools. In selecting tests there should be an emphasis on the instrument's technical qualifications of reliability and validity, and the adequacy of its referencing, either norm or criterion. Practical considerations such as ease of administering, provision of supplementary aids, and economy are also of great importance.

Although standardized tests and other commercially published instruments have traditionally been the prime sources from which test selections have been made, there may be exceptions to that rule. As an extreme example, the primary purposes of the testing program may lead toward the "in house" development of testing instruments based totally on the local educational program. The availability of "objective banks" with sample evaluation items suggests this. A caution should be made that such a testing program has an inherent limitation -- it would be referenced only to itself. The value to a school system in knowing that its pupils are meeting the objectives of its own instructional program is only part of the educational picture. What happens if its instructional program is inadequate to the needs of society outside the school community? it is doubtful that any sophisticated community would allow such a provincial evaluation procedure to develop or to continue without questioning its effectiveness.

SCHEDULING TESTS APPROPRIATELY

Next, decisions have to be made regarding the testing schedule. The particular use of the tests may suggest a certain time of the year as being most appropriate. The administration of tests early in the school year may have advantages of providing the most up-to-date information for use during the current year. With fall testing, the emphasis is usually construed to be the analysis of student needs. With spring testing, the emphasis is often construed to be using the tests for evaluation of teaching. In order for spring results to be used maximally

for students, the individual results must be reported to the schools which the students attend the next fall.

In selecting testing dates, consideration should be given to testing at approximately the same time of the year that the normative data for the tests were gathered. Administering the tests at other times necessitates either the mathematical interpolation of the norms or the use of inappropriate norms. Many test publishers now obtain empirical norms two or more times of the year at each grade level. This practice makes less critical the selection of testing dates that correspond exactly with the dates reported by the test developer.

ADMINISTERING TESTS

Planning a testing program also involves the detailed specifics of administration. Lack of attention to this seemingly minor detail can spell disaster. Even if the best tests have been selected and the most exquisite statistical analyses have been used to present the data in their most alluring form, it is wasted if the test administration was improper. If the results of the program are to be used, the testing must be carried out in such a manner that no one can question the results because of improper administration. This means that in the planning of a testing program, it is necessary to see that adequate time and effort are given to orientation and training of those who give the tests. Systematic procedures must be established to insure that all necessary information is clearly provided, that administration of the test is standard, and that scoring is correct and carefully checked. It is also necessary to see that other details such as physical arrangement, adequate seating, and sufficient proctors are provided as needed. (See Memo No. 4).

PLANNING FOR USE OF TEST RESULTS

Ultimately the success or failure of a testing program depends on its usefulness. The data provided by a testing program are no better than the degree to which they are communicated. The results of a testing program must be communicated in forms that the primary users can understand; therefore, planning a testing program involves careful attention to developing procedures and techniques of communication with these users. This suggests a multiple approach. The expansion of the media field in recent years provides fascinating resources to explore in getting the message across. Care should be taken, however, to assure that those fascinating media resources do not become the focus, obtusating their message.

It is particularly important to assure that at some point personal contact is made in communicating test results. Mass education leads to readily to mass techniques of communication. To make test results real, meaningful, and thereby usable, human interaction is needed. Providing planned situations for the sharing of results; confirming, questioning, hypothesizing, relating, and interpreting are essential. (See Memo No. 6, Chapter III).

PROCESSING AND RECORDING TEST RESULTS

The application of data processing technology to testing programs has brought mixed blessings. The speed of computers and scoring machines dazzle the imagination with almost magical possibilities. But "turn around" time for a teacher to receive the student's test scores lags far behind that of the interested and professionally minded teachers who prefer to score their own tests and have the results ready to use one or two days after administration.

Decision makers need timely information. In planning, careful consideration should be given to reducing to a minimum the time between test-taking and reporting results. Commercial scoring services can usually provide two or three week turn-around time. Institutions having their own processing equipment appear to have the potential of providing a faster service. Careful checking may show, however, that institutional test scoring and processing have a priority far below financial and other functions. Clear and definite commitments to provide timely scoring and reporting of tests should be obtained from whoever provides the service.

The system of processing in most cases must be coordinated with the record keeping system. Test labels and group summaries must be adapted to fit the record forms in use, or the record forms must be adapted to fit the pre-printed and group summaries. (See Memo No. 6).

PERSONNEL NEEDS

A testing program cannot run itself. The quality of the program will reflect the quality of the personnel who implement that program. The background and training of the persons selected to operate the program is likely to influence the direction of its development. A staff dominated by those with backgrounds in pupil personnel work is likely to move a testing program in a direction quite different from a staff dominated by those trained in curriculum or research. All functions of the educational program must be considered as planning progresses.

MEMO NO. 4: SUPERVISING TEST ADMINISTRATION

Gerald S. Hanna

Kansas State University

Among a test director's duties, one of the most worrisome is the supervision of numerous teachers who may be simultaneously administering a standardized test in several locations throughout a district. Numerous pitfalls await the unwary on the path to useful test data. One is haunted with questions. "Can the teachers follow the standardized procedures?" "Will a vice principal decide to have a fire drill during the test?" "Will I know if someone makes major errors in the administration?" "Will Ms. Peach misplace a test booklet again this year?" "Will the teachers rise in revolt if we have a training session again this year?" This memo is devoted to discussing some ways in which the test director can increase the degree to which testing produces dependable, useful information. Attention will be directed to careful, detailed planning for the test administration, enlistment of teacher cooperation, training test administrators, and quality control.

PLAN, PLAN, AND PLAN SOME MORE

Detailed plans for all phases of the test administration are needed several weeks in advance. The surest way to endanger successful administration is to dismiss "mere logistical details" as trivia to be casually delegated to an inexperienced and/or nonprofessional subordinate. There is nothing "mere" about planning to have the right materials in suitable testing environments at the designated times to be administered by competent examiners. Careful attention to such details will pay enormous dividends in difficulties avoided.

The examination date and time in each school (which may or may not be uniform throughout the district) should normally be established several months in advance and should be set only after consultation with appropriate administrators in order to avoid conflicts with exciting athletic activities, field trips, parent days, etc. Having established testing dates, one should decide upon the room or rooms in each school in which the test will be administered. This process usually degenerates to a matter of opting the least objectionable alternative. On the one hand, it would often be desirable for purposes of uniformity and convenience to have many students tested in large rooms. On the other hand, it is usually painfully obvious that the large areas available (e.g., auditoriums, libraries, and cafeterias) were designed for other purposes and seldom provide ideal testing conditions such as generous working space, freedom from distractions and spacious arrangement of students. Moreover, using such rooms for testing tends to interfere with the primary functions that such facilities were designed to fulfill. Then too, in testing younger children it is desirable to examine them in the familiar environments of their own classrooms. After much thought and

negotiation, the decision is often in favor of testing in the individual classrooms.

This necessitates having more examiners than would be necessary for testing larger groups. In school employing a counselor or administrator who would have time to repeatedly administer the test (That will be the day!), only one or two test administrators per school may be needed. But most probably it will be necessary to enlist the assistance of classroom teachers in giving the tests. In spite of the problems this may raise, it often has the valuable benefit of causing the teachers to feel they have a time investment in the testing program; this may increase their propensity to make good use of the testing results.

Having decided where and by whom the test will be administered in each school, it is necessary to plan for the absence of interruptions. Administrators, having many other things on their minds, are sometimes susceptible to forgetting that they have promised not to have a fire drill or intrude via the intercom system with a laudable ping for the PTA membership drive. Frequent reminders help. Also, it is often useful in enlisting the interest and support of the secretary or clerk who is in a position to remind the principal of acts he ought not to commit by way of interruptions.

Another useful precaution is to secure "Testing, please do not disturb" signs to be posted conspicuously on each door of rooms used for the test administration.

Another requirement for planning is the receipt of testing materials and information at each school. This requires the designation of one person in each school to serve as coordinator, being responsible for safely storing the tests, distributing them on the designated days, collecting them again, sending the materials back to the central office (if desired), disseminating directions about the program to the examiners, reporting problems to the test directors, etc. Principals are frequently too busy to give this job the detailed attention it demands. A teacher or counselor who is mutually agreeable to the principal and to the test director is often preferable. In all but the smaller school districts the test director will want to have local school coordinators designated early enough to perform most of the planning functions mentioned above.

The matter of storing test booklets and answer sheets deserves more than a fleeting thought. It is imperative, of course, to protect test security by storing materials in a place where curious students, clerks, custodians, etc., cannot browse and borrow. This point is equally true at both the central district and the local school levels. Similarly, arrangements for transporting material to and from individual schools need to provide for punctual delivery to designated persons, security from loss of materials, and safety from damage.

A locally devised form or checklist is useful in distributing and packaging testing materials. Original and carbon copies of such a form can simultaneously serve (1) as a checklist for the sender to be sure that the intended number of copies of test booklets, answer sheets, manuals, local memoranda, etc., are enclosed, (2) as a packing slip for the receiver against which to "take inventory," and (3) as a central office record of who has what.

Another of the less challenging aspects of supervising a test administration is counting testing materials repeatedly. When materials arrive from the publisher, they should be counted accurately enough to ensure a sufficient supply. When these are packaged for local schools, they should be counted and recounted, preferably by a different person. When they arrive at the school, test booklets and answer sheets should be counted by the local coordinator before storing. When they are distributed by the school coordinator to teachers they should be counted and checked. When they are received by the teacher, they should be counted. Similarly, as the test booklets are sent back, step by step, from the teacher to the district storing facility, they should be counted. Why all this compulsive preoccupation with counting? The answer, of course, is twofold. First, counting test booklets, answer sheets, manuals for administering, etc., as they are successfully distributed from the director's office to the individual teacher assures that clerical mistakes will not result in uncorrected shortages of materials. Being ready to administer a test to thirty students and discovering that one has only twenty-nine test booklets or answer sheets is not only awkward, but inexcusable. Second, counting tests at each stage of distribution and collection makes it relatively easy to pinpoint blame for missing tests. It has been the writer's experience that the ability to isolate blame for missing tests almost always prevents the loss of the tests in the first place.

ENLISTING TEACHER COOPERATION

Teachers are likely to be motivated to prepare carefully for the test administration, attend a training workshop, communicate positive feelings about a testing program, etc., if they see it as helping them. It is therefore, necessary that teachers be involved in the testing program long before the test administration. To the extent that teachers perceive the test as helping them do their job better in ways they believe in, they will be cooperatively disposed. To the extent that they perceive the testing program as something that is started and managed by some "experts" who run the whole show (and that is you!) they are less likely to identify with it and work for its success. It is easy to adopt the latter attitude.

A point of view from cognitive-field theory may be advantageously kept in mind. One's attitudes are a function not of reality but rather of one's perception of reality. Therefore, it isn't enough that the test results could be helpful to teachers; teachers must perceive the

results as being useful if their cooperation is realistically to be expected.

The moral for the test director is to remember that, although the potential utility of the test results may be obvious to him, it may not be at all apparent to many teachers. And in a very real sense, the teachers' perceptions of test utility are more real than ours, because it is the teachers who have the most opportunity to use the test results. "If test results actually do not influence educational decisions, then it is foolish of the school to waste the time and money involved" (Nunnally, 1972, p. 520).

TRAINING THE TEST ADMINISTRATORS

If the decision is made to have classroom teachers, rather than persons more specialized in measurement administer the test, it is desirable to provide them with an orientation or training session. Indeed insofar as the professional preparation of teachers often leaves something to be desired concerning their measurement competencies (Mayo, 1967), it may be essential to provide some training. "Although in ordinary testing programs, classroom teachers can administer the group tests if they are given proper orientation and supervision...it is not sound policy to take the examiners' competence for granted," (Stanley and Hopkins, 1972, p. 429).

The content of the training or orientation should provide the examiners with detailed knowledge of how to administer the test and with practice in doing so. Emphasis may be placed on the need to study the part of the test manual that details the administration procedures before the testing session, the need to keep accurate time to the nearest second, how to keep time, the need to speak slowly and distinctly, establishing rapport, motivating examinees to do their best, efficiently distributing and collecting materials, the need for rigorous adherence to standardized procedures, preparing students for testing, answering (or not answering) student questions, avoiding disciplinary disruptions, avoiding cheating, and reporting events that might tend to invalidate test scores. The orientation should also be designed to familiarize examiners with supplementary directions concerning such things as counting supplies, having extra pencils and erasers available and alphabetizing the answer sheets. When planning an orientation or training program, reference to standard texts (e.g., Clemans, 1971, pp. 196-199; Cronbach, 1970, pp. 50-78; Gronlund, 1971, pp. 348-352; Thorndike and Hagen, 1969, pp. 539-544) would assist in identifying points to be included.

In larger school districts, it will often be impractical for the test director himself to conduct training sessions in each school. It may be necessary to enlist the help of local coordinators or others, e.g., counselors, to provide the number of training sessions required. Persons conducting the local training should, of course, be selected to assure some professional preparation in educational measurement.

In addition, the test director normally will want to provide them with some instruction in training test examiners. Observing and/or assisting the test director in the conduct of training sessions for teachers may also be useful.

The training or orientation of examiners could be presented by means of several media. Most common is a meeting or workshop conducted by the test director and/or school coordinator using interrupted lecture and/or demonstration. A video-tape or motion picture could be used to good advantage in providing a demonstration of proper test administration and for presenting problems for discussion. Such a "canned" demonstration can excel a mock live demonstration both in realism and in control over selection of points to be illustrated, such as questions asked by examinees. A module or series of modules could be built to provide minimum essential competencies for test administrators. A particular advantage of this approach would be the possibility of individual examiners avoiding unneeded instruction by demonstrating competence on pretests. Of course, a combination of the above and/or other media might be selected to provide an instructive and interesting orientation.

Some schools have attempted to preclude the need for more than one examiner per school by the use of the school intercom system or tape recorders for test administration. However, this introduces problems of synchronization and problems of answering allowable student questions before the test begins. Although in this pattern the teacher proctors rather than administers the test in his classroom, the kind of proctoring required would demand most of the competencies mentioned above. There is some question as to whether more problems are solved than created by this kind of arrangement.

QUALITY CONTROL

In this final section a brief examination will be made of the need for the test director to provide himself with feedback, to correct deficient practices, to know when to void examination results, and to arrange for make-up testing.

Concerning the need for feedback, it is necessary for the test director to have sufficient rapport, esprit de corps, and nonthreatening relationships with school coordinators and teachers for the latter to feel free and unafraid to report things that went wrong. Indeed, it is wise for the test director to actively look for problems, but in doing so, he should strive to communicate that his objective is improvement of the program, not criticism of personnel. In supervising one's first major test administration, it would be wise to serve as a coordinator once or twice to assist in identifying and correcting problems at that level. Subsequently, the test director could be the "substitute teacher" a few times to retain a clear image of the test administration on that level.

Once the procedures are operative, one should not consider the job finished, but should remain forever sensitive to feedback which can be used to correct deficiencies in the administration. For example, you may learn that in spite of the excellence of your training session, one teacher has been assuming that her stop watch is self-winding, another has been telling students not to guess on questions the answers of which they do not know because it is dishonest and immoral to guess, and yet another well meaning soul is giving her students an extra five minutes because they are slow learners. Unnerving revelations such as these are more likely to be made if you are visiting with teachers in their faculty rooms than if you are securely insulated in your office. The solution for future administrations to some of these kinds of non-standardized practices may be to stress the relevant points in the orientation. (But be assured that new ingenious errors will be generated the next time the test is administered.) The solution to other problems may be to work with individual teachers to increase the extent to which they appreciate the need to adhere to standard procedures. And the solution to a very few cases may be to recognize that one or two of the teachers may be functionally unable to follow directions; hence, they should not administer standardized tests.

In the process of obtaining feedback with a view to subsequent improvement, another equally important purpose is served --- determining which tests, if any, should be voided because of examiner error, invalidating interruptions, etc. A feeling of professional partnership among the test director, school coordinator, and teachers will increase the likelihood of testing irregularities being brought to your attention. Whenever the circumstances of administration cast doubt upon the validity of a test, the results should be voided. Compulsive desire to have a score for every student should never obscure the truth that no information is preferable to misinformation. One should be willing to err on the side of caution by adopting the practice of voiding a score whenever its validity is in doubt.

A make-up test administration should normally be scheduled for students who were absent when the test was administered and for those few whose scores have been voided. An alternate form should normally be used. It is regrettably easy to let all of the elaborate precautions against loss of tests, administration by unqualified or nonoriented persons, and deterioration of quality control lapse and to be careless in supervising the make-up administrations. In point of fact, of course, there is just as much need for professional thoroughness in supervising make-up tests as there is for the original test.

This memo could be summarized with the caveat that if anything can go wrong with the test administration, it probably will somewhere in the district. The function of the test director is to see to it that (1) relatively little can go astray and (2) he knows about it when something does go wrong.

(Note: Some additional suggestions for test administration are contained in Chapter VI, "External testing Programs". Many of these suggestions are also useful for the administration of any standardized test.)

REFERENCES

- Clemens, William V., Test Administration in Thorndike, Robert L. (Editor), Educational Measurement, Second Edition. Washington, D.C.: American Council on Education, 1971.
- Cronbach, Lee J., Essentials of Psychological Testing, Third Edition, New York: Harper and Row, Publishers, 1970.
- Gronlund, Norman E., Measurement and Evaluation in Teaching, Second Edition, New York: The Macmillan Company, 1971.
- Mayo, Samuel T., Pre-Service Preparation of Teachers in Educational Measurement, Chicago: Loyola University, 1967.
- Nunnally, J. C., Educational Measurement and Evaluation, Second Edition, New York: McGraw Hill Book Company, 1972.
- Stanley, Julian C. and Kenneth D. Hopkins, Educational and Psychological Measurement and Evaluation, Englewood Cliffs: Prentice-Hall, Inc., 1972.
- Thorndike, Robert L. and Elizabeth Hagen, Measurement and Evaluation in Psychology and Education, Third Edition, New York: John Wiley and Sons, Inc., 1969.

MEMO NO. 5: OPERATING A SCORING SERVICE

by*

Richard L. Poole, The Ohio State University
 Jack Menne, Iowa State University
 Annie W. Ward, The University of South Florida

An essential element of a school testing program is the scoring of tests. In regard to this a test director has three choices open to him:

1. He can decide that all tests will be hand scored, which usually means being scored by the classroom teacher.
2. He can contract with the test publisher or a commercial agency to do the required work.
3. He can set up a central scoring service for all or some of the tests used in the school system.

In view of these choices let us now consider some of their possible advantages and disadvantages.

SCORING TESTS BY HAND

This choice seems to work best for tests used in the primary grades. Many elementary school teachers report that they like to see how the students responded to the test questions, and they also like the acquisition of immediate student results. Also, some tests for young children are only available in a hand scorable edition. If a hand scoring approach is selected, training sessions should be planned and held to provide teachers with the opportunity to develop efficient and consistent scoring procedures.

At other grade levels within the school, hand scoring is not as well received by teachers. The possible reasons for this are unclear but may include the following: (1) It is known that most standardized tests for grades four and above are available with machine scorable answer sheets. (2) The task of counting "filled in" or "blackened" spaces on answer sheets does not appeal to teachers. (3) The implicit expectation to do this is missing.

It is recommended that teachers be used in the hand scoring of tests only for the primary grades, if at all, and that they not be expected or required to convert the acquired raw scores to other types of derived scores.

*Many helpful comments and suggestions were also provided by Edward Caldwell, Director of Evaluation Services, University of South Florida.

The raw scores may be reported to the testing office in a variety of ways. The simplest is for teachers to send the entire test booklet for each student. This makes it possible to spot check the scoring as well as to process the scores. A second choice is for teachers to tear off the first or last page of the test booklets, on which they have written the students' name and raw scores, and send these sheets to the testing office. A third possibility is for teachers to prepare a list report of students' names and raw scores.

The conversion of raw scores to normed scores and preparation of all reports should be done centrally. If data processing equipment is available, the task is very simple. It simply requires keypunching and computer processing. If there is no such equipment available, trained clerks should do these tasks under the test director's supervision.

SCORING TESTS BY MEANS OF CONTRACTED SERVICES

This choice appears to be most appropriate for small systems and for school systems with limited resources or a limited testing program. Many test publishers, some commercial scoring services, and some universities and colleges provide scoring services which can produce various types of test related information (computer output) with a corresponding variable scale of costs. The unique feature of most publisher affiliated scoring services is the wide range of derived scores which can be produced at a relatively small cost per pupil. Yet, the educational or instructional relevance of all these derived courses is sometimes hard to see. The disadvantage of this type of service is usually the time lag between test administration and the return of student and test data to the school.

The university or commercial scoring service may not be able to match the array of scores and services available from a test publisher, but often can supply limited but perhaps more practical and less expensive service in a shorter amount of time.

In any case, the service and scores provided by any of these agencies is probably much lower than those the school system would secure if it tried to provide similar services locally. Also, it may be the only option for small school systems. (See section on costs below.)

SCORING TESTS BY MEANS OF A SYSTEM SCORING SERVICE

This choice is open to the large school system which has a data-based decision-making policy with regard to schools and students, as well as a commitment to an extensive school testing program including the processing and analysis of teacher-made classroom tests. This approach requires necessary resources of people, space and a computer facility which has the capability for expansion. A locally operated school scoring service can provide many services. It can score, norm and otherwise

process most standardized tests; it can score, norm and analyze (item analysis and statistical analysis) teacher-made tests; it can process, tabulate and analyze many evaluation instruments such as questionnaires and rating scales; it can provide customized output for explicit school purposes; it can provide the data to make decisions by design not by default, and it can do all this, in all probability, with a very short lag time or turn-around time.

In addition to these services, it can also provide for the development of data banks and the printing or reproduction of tests by the computer. The former activity would make it possible to follow individuals and groups for long periods of time, and the latter activity would make it possible to print tests accurately. These considerations are so important that any school system which has the capability will almost certainly opt for operating its own scoring service. The remainder of this memo will discuss the procedures for setting up the local scoring service, operating it effectively, and optimizing its use.

ORGANIZATION OF SCORING SERVICES

Test scoring and related activities are service activities. A more appropriate name than "test scoring" may be "computer-based instructional support services." These services must be of a high professional quality and must be prompt; if they are not, the purpose of the service will not be accomplished, volume will go down, and unit costs will increase.

COSTS

Before discussing procedures for setting up scoring services, let's first talk about costs. Let's assume for this discussion the scoring service is paying \$1000 per month for machine rental and maintenance. If the scoring service processes 12,000 sheets per year, the machine cost alone would be \$1 per sheet! At 600,000 per year the machine cost would be 2¢ per sheet. Machine cost is essentially a fixed cost; volume brings the cost per test or sheet down.

Similarly, the cost of the clerk to operate the service will be essentially a fixed cost. The scoring clerk could be assigned other duties; however, installations which have tried to operate with part-time help or clerks with multiple duties have often found that the quality and promptness of the service is not ideal and user volume drops off.

Some installations have been able to operate with machine and clerk processing costs of about 3¢ by getting the volume above 1/2 million sheets per year. There are installations operating with less than 100,000 sheets per year which could purchase their processing (and perhaps even improve service) at less cost than those incurred in the present low volume installations. These processing costs should be studied in planning a service and should be monitored thereafter. Remember, too, that besides processing costs, there are costs for answer sheets, computer time, space, etc.

SPACE REQUIREMENTS

A scoring service requires a great deal of space. Space must be provided for these activities:

1. Receipt, sorting and cleaning of answer sheets.
2. Storage of used answer sheets and cards for some installations. Answer sheets must be stored until they are scored, and, after scoring, they must be kept until the entire processing job is completed and verified. If a card output is used, new cards must be stored near the scoring machine, and used cards must be kept until the processing has been completed and verified and the results have been put on tape.
3. Operation of auxilliary equipment, depending on the type of scoring equipment used, should be located in an air-conditioned area which can be insulated for noise. At least one large table or counter is essential for temporary stacking of papers and cards.

Test scoring is at times located with and even a part of the general data processing organization. Since the first priority for this operation is usually payroll (for which we are very sympathetic!) and other financial operations, the test scoring operation may be relegated to the lowest priority, and may be done at those times when there is nothing else to do. Some years ago it made good economic sense to coalesce all computer related activities; now however, the mini-computer has dramatically changed this economic situation. A test scoring operation ideally should not be a part of a general data processing service and should not even rely on such a service for computer support. A mini-computer can provide the necessary processing right in the testing office.

If a central computer must be used, it would be convenient to have the office of the test director, the school scoring service, and the computing facility in close proximity to expedite their synergetic nature. However, if the computer center is not near the test director's office and a decision has to be made between locating the scoring service near the test director versus locating it near the computer, the choice should be in favor of the test director. Sorting and preparing the test papers is a necessary part of the operation, as is checking the output of the scoring center. Ethically and practically, the test director must involve himself in both these activities. If the scoring service is located at some distance from his office, he will spend a great deal of his time there, effectually cutting himself off from other activities which need his attention.

Such a separation between the scoring service and computing center does necessitate the transportation of computer cards to the computing center, but that is the lesser of the two evils.

EQUIPMENT

Equipment for scoring tests and processing test data has changed dramatically over the years. The original scoring machines like the IBM 805 operated through electrical contact to the mark on the sheet, hence the need for pencils with special lead. Machines next developed like the IBM 1230 and the DIGITEK (now Optical Scanning Corp.) models which operate by reflected light and are properly termed "mark sense" readers in that they "sense the presence of marks."

The older mark sense readers are available as test scoring machines. That is, they have sufficient internal logic capabilities to sum the number of correct responses into a score and print this score on the answer sheet. They also punch cards or tape for computer input. Newer equipment includes the following components:

- a) document reader for reading marks.
- b) mini-computer
- c) a printer with the capability of listing students by name and/or number with score or scores.

One manufacturer has combined these three elements into a package so, in a sense, it markets a "test scoring machine" though the scores are not printed on the original answer sheet. Just recently scoring machines which score test booklets (not just answer sheets) have become available. These may be used for tests previously hand-scored. These readers process paper sheets at least 8 1/2" x 11". A number of readers that process smaller materials -- usually card stock like 3 1/2" x 6" are available but this size material generally meets with user resistance and does not have the versatility of 8 1/2" x 11" paper. In using the equipment to process surveys, student ratings of teachers, questionnaires, etc.

The test director has a choice of purchasing or renting a self-contained, all-function unit or assembling a configuration of equipment from a variety of sources. The advantages of assembling one's own configuration are the probable savings on equipment costs and greater versatility in applications. The disadvantage is the need to develop one's own "software" that is, the program packages to carry out the desired processing.

Some sources of the document readers and/or "test scoring machines" are listed below:

Optical Scanning Corporation
Newtown, Pennsylvania 18940

IBM
Endicott, New York

Decision
155 Main Street
Danbury, Connecticut 06810

National Computer Systems
4401 W. 76 Street
Minneapolis, Minnesota 55435

The list is suggestive. There may be other manufacturers who offer similar or more advanced equipment by the time this is read. The test director should be prepared to investigate the market rather exhaustively before making a commitment to either purchase or lease equipment. Indeed, even the decision as to lease or purchase needs careful consideration. One should talk to other test directors who have similar kinds of testing programs in order to explore the various possibilities.

ANSWER SHEETS

One persistent problem in a scoring service is answer sheets. Most machines will read only the answer sheets printed to the specification of the manufacturer --- and manufacturers do not have common specifications. In some cases, a manufacturer with more than one model of machine actually has different answer sheet specifications for each machine. Also, some of the printing specifications are so intolerant that local print shops cannot be utilized. As the cost of answer sheets will be of the same general magnitude as the equipment costs in a large volume service, it is necessary to consider carefully answer sheet specifications. Some answer sheets can be printed locally and one vendor produces a reader which will process up to four different types of answer sheets. Another vendor now has a reader which reads test booklets. It is clear that the type and numbers of documents which can be read must be a consideration in the selection of a document reader.

PROCESSING

The processing of test papers involves these steps:

1. Reception: This includes the physical acceptance of papers, identification of the desired scoring and processing services, and assignment of priority. Most scoring services have developed a "Request form" or "Request record" which is completed by the client and sent with the papers to the scoring center.

Some scoring services have found a daily log form simpler to use, particularly with clients who bring in their papers in person. This form is also useful for system-wide testing programs. The log form is completed by the receptionist when the papers are brought in. Whether a "Request Form" or "Daily Log" is used, it should cover the available services and be arranged so that it is easy to summarize the services which have been provided.

2. Preparation: Answer sheets must be sorted into scoring groups along with the appropriate key for each. For system-wide testing programs, it is helpful to prepare and check in advance the keys for all answer sheets to be scored. Each key is carefully marked and put into a folder. Then, when the tests are sorted out, the folder is laid on top of the stack of papers for scoring. After the papers have been scored the folder with the key is returned to the file. For the

purpose of security, it may be desirable to have a burglar proof file with a combination lock. Teacher-made tests must have their own teacher-made keys which are well-marked and simply laid on top of the group of papers to be scored. As answer sheets are sorted, attention must be given to their condition. Scoring machines can be programmed to handle a great many problems, but it is still necessary to give the papers a personal inspection because of specific machine operating characteristics. For example, a machine can be programmed to read the darker of two marks, or it may be set to reject papers which contain double responses. But in this latter situation, what should be done with the paper? Should it be returned to the teacher for hand marking, or should the machine operator erase one of the two marks? Sometimes a writing tool other than a pencil is used, and if scored would in all probability yield defective results; hence the paper must be pulled and re-marked or marked over with the correct type of pencil. Also, stray marks in the area of the timing marks, as well as papers which have dog-eared corners, folds, wrinkles or staples can create serious mechanical problems. Obviously, the best solution is to spend time in training test administrators and scoring personnel in techniques of how to handle and avoid these hazards. However, a check must always be made to be certain there are no difficulties before answer sheets are loaded on the machine.

3. Scoring: The operation of the scoring machine requires first setting up the machine properly. Then answer sheets are run through. Specific instructions for using the scoring equipment are available from the manufacturer's representative and cannot be covered in this memo. It is the responsibility of the local staff to see that the correct key is used, that the dials are set correctly, and that the card format conforms to the requirement of the computer programs to be used. It is also important that the accuracy of the scoring be insured by periodical checks of the scoring key used and the actual scores.

4. Computer Programs: The processing of data requires a great deal of effort which is best handled by a computer pre-developed programs. For classroom tests there should be a packaged program which includes as a minimum the following information as output: A frequency distribution of raw-scores; a conversion of each raw score to the derived scores of percentiles and T-scores; item analysis data; reliability data; a classroom roster by raw-score and T's or percentiles; a listing of errors by class and by student.

Programs for standardized tests must include conversion to national and local norms, rosters of all scores; class, school, and district summaries and labels for recording results for individual students on their records. The printing of profiles and other special reports may be added if desired.

COMMUNICATION

There must be a great deal of communication with the user if the test-scoring service is to be maximally effective.

For even routine scoring jobs, such as the scoring and reporting of the system-wide standardized testing program, communication with teachers about the handling of answer sheets can prevent many headaches in the scoring room. And, when the results are returned to the schools, it is essential that everyone know how to interpret the reported data. Written materials help, but it is also useful to conduct seminars and personal visits with users to help them read and understand the reports. When the scoring service is to be used for classroom tests, the communication function becomes even more important. Not only do most teachers need some help in the area of planning and writing good tests, they also need specific help in setting up tests in a machine scorable format, writing instructions to students, handling the answer sheets, and preparing the key. In order to use the test-scoring service optimally, they also need to know what scoring and programming options are available and what kind of information each option provides. When the scoring is completed, they also may need help in interpreting the information and in using it to improve their tests and instruction.

Scoring services can and should be used for processing of other evaluative data. This may require the development of new instruments or it may simply involve the conversion of existing instruments for use with an answer sheet which is machine processable. In general, instrument conversion necessitates a change in the directions, so that the respondent writes his responses on an answer sheet rather than on the instrument itself. There is the possibility that such a service might also mean assisting in the development or formulation of such an answer sheet. If budgetary considerations are of little concern, then the consumer may have a customized form drawn up and printed by the manufacturer. However, for low budget projects, it would be wise for the consumer to explore with the scoring service staff the possibility of using a standard answer sheet. Over-printing one of the standard forms may also be done with a ditto machine (but not a mimeograph). Surveys, rating scales, questionnaires and many other types of instruments may be designed for use with separate answer sheets, and the saving in processing time and their greater flexibility more than justify the time which must be spent in writing items and precoding possible responses.

BUILDING A DATA BANK AND WRITING TESTS BY COMPUTER

There are many reasons for building a data bank of student information, particularly if there is a regular program of system-wide testing at certain grade levels. With a data bank it is possible to provide these additional services:

1. When students move from one level of education to another,

pertinent test data may be reported to the receiving school very easily.

2. When information is needed to evaluate special programs for which the regular tests are valid, the evaluative data may often be pulled from the data bank without additional special testing.
3. It is possible to accumulate data and follow individuals and groups through many successive years and experiences.
4. The testing program itself may be validated over a period of time.

In order to build a data bank it is necessary to determine first what data will go into it. Clearly, if test data are to be included in it, then a complete title of the instrument should be included along with any raw and possibly derived scores. In addition, if report cards are computer-generated, then they too may be included.

In order to make the data bank as utilitarian as possible it is essential to assign and maintain a unique student number for addressing each student file. Eventually all children may be assigned an identification number at birth, but for the moment this is not the case. At some high schools and colleges the social security number is used for this purpose, but most elementary pupils do not have such a number, so a system has to be developed for assigning numbers to existing students as well as transfer students. One way that has worked successfully is to have the number assigned at the first school which the student attends. The first five digits form an "accession" number and the last four digits identify the school. So long as the student stays in the school system his accession number remains the same. If he leaves the system before graduation his number becomes available for a transfer student, but his number may or may not fit into the system of the new school system. Such issues as these and many others await those who aspire to build a data bank related to student data.

Another aspect of the data bank is an item or test bank; the test items or even complete tests are written into the computer for subsequent retrieval. The identifying codes are assigned to an item, group of items, or even a complete test. Depending on the nature of the retrieval program it is possible to scramble the test item options or the items in a complete test. With the advent of individualized and modularized instruction, the capability of printing or "writing" one test at a time is no longer a utopian phenomenon.

To summarize: operating a school-related scoring service affords school personnel many benefits. These benefits are not without costs; however, the costs may be justified in terms of service, and this service can and should be more than the routine processing of answer sheets for a systemwide standardized testing program. A scoring service pays by serving as the nucleus of an information system which can quickly and accurately provide needed student or school program information.

CHAPTER III

USING RESULTS OF STANDARDIZED TESTS

Margaret E. Backman
Research Director
ICD Rehabilitation and Research Center
New York, N. Y.

In many school systems standardized testing has become an institutionalized ritual that is followed (or endured, as the case may be) with little enthusiasm. Yet the information that can be gained from testing should be a necessary and welcome part of the instructional and learning experience.

Projected uses of test results, as they appear in proposals for testing programs and measurement publications can make valuable contributions to a school program. (See Introduction, Chapter I). Why then are such objectives only partially realized in most school systems?

For one thing, the responsibilities for testing are often assigned to guidance counselors, vice principals, or willing teachers. Limitations of time, money, and in some cases technical know-how tend to make the administration of the tests the major focus; after that, there is little in the way of follow-up or reporting of results. No wonder students and teachers come to feel that standardized testing is an annoying interruption in the school routine.

The responsibilities of planning, managing, and providing professional-technical services constitute a full-time job, and should be the sole responsibilities of one person, the testing director. In large school systems, the testing director needs the assistance of a staff or at least people in each school who have time allocated to work on the testing problem. In addition, for the testing program to be effective it must have the support of all concerned groups: faculty, students, administrators, educational and guidance specialists, parents, and in many cases the tax-paying citizens. The role of the testing director becomes somewhat complicated, however, because the people comprising these groups have diverse needs, interests, and levels of sophistication.

The testing director should keep in mind that there are those who really need and can use test results, those who accept testing without question, those who go along but with reservations, and, last but far from least, those who are vehemently anti-testing. Those who are strongly in favor of testing are often very supportive of the program, but can present a problem if they give more importance to test results than the results merit. Those who just go along with the testing, may be so uninvolved that they will never look at the test results, and will thus miss out on much

useful information. Those who strongly resist testing in any form are often the greatest obvious source of frustration.

It can be discouraging for the testing director when his efforts are met with resistance or, at best, apathy and disinterest. The apathy may come from those who do not understand how to interpret the results of testing, and thus feel they get little from the time and effort required. Some who resist testing do so because they feel threatened by what the tests might reveal or how the results will be used. And others, probably the most vocal of the anti-testing people, feel that tests are dehumanizing, not valid, and biased.

In some instances the criticisms may be valid; in others they may be based upon misinformation. The issues raised by those resistant to testing, however, should not be ignored. Criticism can serve as a needed check, and make us question what we are doing instead of blindly doing what has been done before.

"TESTS ARE DEHUMANIZING"

The criticism that "tests are dehumanizing" comes about partly because tests are seen as reducing people's complex make-up to a single number. In responding to this criticism, the testing director should stress that test results should only be used to supplement other information; the test scores are only an estimate of a person's ability in a given situation.

The testing director should also point out that some people, such as the guidance counselor, need as much information as they can get to advise students. Tests can provide information about the growth of individuals; they can be used in the identification of the gifted and the so-called underachievers. If it were not for objectively scored tests, some students who have intellectual talents might be overlooked: the unruly boy, the pretty girl, the minority student.

There are those that feel that testing is dehumanizing because people are compared to one another on the basis of test scores. This, they feel, creates unwanted competition, and for some, a recurring sense of frustration and failure. The movement toward testing for mastery is sometimes seen as a more acceptable practice (see Bloom, Hastings, and Madaus, 1971). Unfortunately, there are few published mastery tests, and many problems in the development of such tests have yet to be worked out. The testing director should assist interested teachers in developing methods of evaluation that are more related to measuring mastery of material than methods currently used.

In part the resistance to standardized testing comes from a misunderstanding of the term *standardization*. Many believe that it is the students themselves who are being standardized; that they are somehow being shaped

into a common mold--*homogenized*, if you will! The testing director should explain that standardization refers to the procedures used in constructing and administering a test. It is the conditions under which a test is given that should be standard or the same. If this is not so, then the scores will have little meaning. It is very difficult to make any sense out of scores that have been obtained from students who have received different kinds of directions, different amounts of time for work, and different kinds of help from the teacher during the testing period.

"TESTS AREN'T VALID"

To the criticism "tests aren't valid," you might ask, "Valid for what?" It is important that users know why particular tests have been selected and what they are intended to measure or predict. Whether or not a test is valid depends upon the particular uses it is intended to serve. Is it going to be used to predict later success, to place students in certain classes, or to measure achievement in a particular subject? A test can be valid for one purpose and not for another.

Also, the glowing claims presented in promotional literature and test manuals are not always borne out by the statistics presented in the manuals themselves. The testing director should evaluate all tests thoroughly to be sure that they will in fact serve the needs of the particular school system. Are the content and processes measured similar to those emphasized in the school curricula? Do the statistics presented in the manual support the kind of decisions that will be made on the basis of the test results, e.g., placement in particular courses? Does past experience with the test, or studies conducted in the school system support continuing the use of the test?

"TESTS ARE BIASED"

The criticism that "tests are biased" assumes that tests are not adequately measuring or predicting the abilities of particular groups. When attempting to respond to criticisms of test bias, the testing director can become involved in an endless, heated controversy. There is no easy answer to what to do in such circumstances. The only solution is to deal with the issues as openly and honestly as possible. Explain the IQ tests, as well as achievement tests, are primarily measures of developed ability at a given point in time. A person's IQ score can change. Controversies can frequently be averted by having representatives of concerned groups help in the selection of tests whose content is appropriate to the needs of the school and community.

A related criticism is that tests are standardized on white, middle-class groups and, therefore, not representative of the abilities of minority groups. It is true that when tests are standardized on large, representative populations in the U.S., the norms are primarily descriptive of the white population. This is so even when minority groups are represented

because the white population constitutes such a large majority in the country. Because norms may not reflect the abilities of particular groups, it is important that the testing director develop local norms, and, if possible, get from the publisher information about the performance of similar groups.

Attitudes like those just discussed probably will have been formed long before the testing director came on the scene. But the testing director can affect these attitudes through good planning and reporting of results. Most importantly, the channels of communication must be kept open.

Thus, the testing director must come to see himself as a public relations expert. This does not imply trying to sell a testing program. What it does mean is that the testing director's role is to help in the development of a sound, valuable testing program, and to see that those concerned with the results understand why the testing is being done and how the results can be of use to them.

In this chapter, three memos deal with problems of interpreting and using test scores for various purposes. The first memo combines guidance and instructional uses, and provides many suggestions for reporting scores to students and their parents. The second memo deals with problems associated with the evaluation of programs and the final memo deals briefly with cautions in the use of test results in the evaluation of teaching personnel.

MEMO NO. 6:
REPORTING TEST RESULTS FOR GUIDANCE AND INSTRUCTIONAL PURPOSES

Margaret E. Backman*

In the planning of testing programs, care must always be taken to insure that the results will eventually be used. Most importantly, the testing program must seem meaningful to those concerned. Procedures should be established for reporting results and the results should be presented in a comprehensible manner.

Following are some specific steps for improving the use of test results in a school system.

Involve concerned persons as early as possible in the planning of the testing program. Concerned persons include not only school personnel, but students, parents, and other tax-paying citizens as well. Such involvement can be brought about through an advisory committee made up of representatives of the concerned groups.

Record results systematically in student folders. Pressure sensitive labels available through scoring services are convenient and efficient; they save time and also have been found to minimize clerical errors. Unfortunately, some scoring services only provide grade equivalent scores, and these, as will be discussed later, can lead to misunderstandings.

Record forms should be set up so the progress of individual students can be followed from year to year. It is best to record achievement, general aptitude, and special aptitude test scores separately. Verbal interpretations of test performance should also be included in student folders, so those not trained in measurement and statistics can easily interpret the test scores.

See that results are presented to individuals and groups according to their needs and levels of understanding. In a large school system, guidance counselors or teachers may be the ones to hold conferences with parents and students. In such cases, the testing director should supply the necessary information in an easily interpretable format. In addition, the testing director should also schedule meetings to help school personnel interpret the test data.

Conferences make it possible for individuals to ask questions and can reduce misunderstandings and misinformation. The personal contact counteracts some of the "dehumanizing" aspects of standardized testing; it can also reduce resistance to the testing program and improve relations with the school system as well.

50

* Some material in this memo was prepared by Anthony J. Allen, U of Rhode Island. Dr. Allen's material is indicated by footnotes.

Reporting results can be difficult, particularly when the "reporter" is in the position of the servant who brings the bad news. In such circumstances, possible steps for remediation should be discussed. For example, a student can be told how to study more effectively; a parent can be given suggestions on ways to help the child outside the classroom; the teacher can be referred to special instructional aids and given supportive assistance from specialists; the PTA can be informed about plans for changes in the curriculum or in teaching methods.

Use non-technical language whenever possible. Instead of presenting test scores *per se*, it is often more meaningful and less confusing to use verbal descriptions of performance.

The following examples represent verbal classifications that can be incorporated into graphs, tables, and written reports.

<u>Percentile</u>	<u>Verbal</u>	<u>Verbal</u>
91st and up	Exceptional understanding of material	Much above average
71st-90th	Good understanding of material	Above average
30th-70th	Average understanding of material	Average
10th-29th	Some difficulty in understanding material	Below average
9th and below	Great difficulty in understanding material	Much below average

The report reproduced in Figure 1 represents the kind of written presentation that can be given to parents and teachers. A recent study found that teachers prefer to receive test reports in a narrative format instead of the traditional listing of student's percentile ranks and grade-equivalent scores (Mathews, 1973). In this example, the report is intended to accompany the histogram in Figure 2.

The attached chart shows Veronica's performance on six tests of scholastic achievement that we gave to her class this year. The six columns refer to her performance on the tests. In all these tests she is being compared with children at her grade level across the country.

As no test is a perfect measure of a student's achievement or abilities, Veronica's test performance has been plotted in the form of a band. The score that represents her actual ability on a test is assumed to fall within the range of scores covered by the band.

According to the information from these tests, Veronica's overall scholastic ability is average or somewhat above average. Veronica has done quite well in Arithmetic Concepts, which measures primarily arithmetical reasoning and understanding. Her Reading Comprehension is also quite good.

Veronica's performance in Spelling and Arithmetic Computation however is somewhat lower than her performance in other areas. She will need extra attention to these skills here at school; however, additional effort on her part might also be effective in bringing these skills up to a higher level.

Figure 1. Report on Student's Performance On Achievement Test Battery.

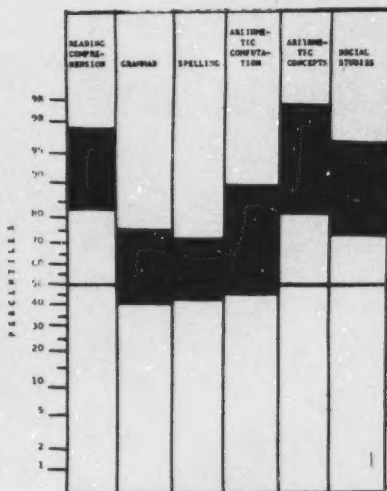
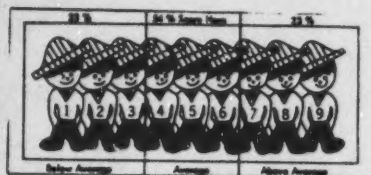


Figure 2. Performance of a student on six tests from an achievement test battery. Bands are based on percentiles corresponding to plus and minus 2 standard errors of measurement from the obtained score on each test.

Another reporting form is presented in Figure 3. This report is addressed to elementary school pupils and presents the test results in a very easy to understand format.

Dear Pupil:

Meet "Stan Stanine" who wants to be a friend of yours. You see nine pictures of him in the chart below. Remember the tests you took this year in the country-wide testing program? Well, Stan wants to help you understand the test results. The more you know about yourself—your strengths and your weaknesses—the better you can plan your school work with your teachers and your parents.



Stan Stanine says:

"The scores on the tests are recorded in terms of stanines. A stanine is a measure used to compare your performance with that of the average pupil in your grade in Volusia County Schools.

Stanine scores range from a low of one (1) to a high of nine (9), with the average being five (5). Now, to show you what this means. Let's say your score on the Spelling Test is in stanine 8. Color my picture with the number 8 on it. What does a stanine of 8 in spelling say? It says that your ability shown in spelling is above that of the average pupil in the county in your grade.

Figure 3. Individual Pupil Report
Designed by LeRosa Smith, Volusia Co. (Fla.)
Reproduced with permission of the author.

Designed by LeRosa Smith, Volusia Co. (Fla.)
Reproduced with permission of the author.

Present results in the form of graphs and tables. Keep graphs and tables as simple as possible, and be sure to label them completely so they can be readily understood. Graphs or histograms describing a student's performance can be given to the student and to his parents during a conference (see Figures 2 & 3). It is good to let people take some material home for future reference. To expedite the drawing of graphs for individuals, reproduce copies of the basic axes with necessary explanatory information; then fill in the data as needed.

When making presentations to groups, large graphs and tables are impressive and, if well prepared, can be very effective. For example, comparisons of test performance can be made between the school system in question and other systems for which norms are available, between different schools within the school system, between students in different grades, for a given class of students as they progress through school, and for the school system over a period of time.

Teachers like to compare the performance of their students with that of other students. A useful tool for each classroom teacher is the class profile which shows, through the plotted line format, the means of that class on each of the subtests in a standardized achievement test; the profile can also show the performance of other groups from the school system or within the state if such data are available.*

The test director should present results in a way that does not put individual teachers on the line. Depending upon the particular school situation, the extent of tracking etc., it might be advisable to condense the scores of all students in a given grade, instead of examining each class as a unit. Similarly, scores of all students in language skills, for example, can be presented to the language arts teachers without reference to particular teachers.*

If the test director is working with a group, he can enhance his presentation of test results by using overhead transparencies of the profiles with overlays for the various norm group profiles.*

See Thorndike and Hagen (1969, pp. 550-555) for examples of graphs applicable for school testing programs, and see Wesman (1966) for instructions on the presentation of data in tabular form.

Report test scores in the form of bands, based on the standard error of measurement. Instead of reporting just a single score for an individual, report a range of scores representing the limits within which a person's "true" score probably lies. Such a range or band reminds us that test scores are only estimates of ability. If a test were to be given again and again, a person's score might fluctuate from testing to testing because of chance factors, i.e., factors not actually related to the ability we wish to measure. Thus, the score representing a student's actual ability could be somewhat higher or lower than that originally obtained.

The upper and lower limits of the bands are based on the statistic called the standard error of measurement (SE). The standard error of measurement is usually reported in the test manual; if not, it can be computed from the reliability coefficient; $SE = \frac{s}{\sqrt{r_{xx}}}$, where s is the standard deviation of the test scores and r_{xx} the reliability coefficient.

The limits of the bands are customarily obtained by adding 1 1/2 or 2 SE to the raw (obtained) score, and subtracting 1 1/2 or 2 SE. For example, if the test manual reports that the SE for a given test is equal to 3 points, we could compute the band around individuals' raw scores by

* Paragraph contributed by Anthony J. Allen.

adding 6 points (2 SE) to each raw score, and subtracting 6 points from each raw score. Thus, for a person receiving a raw score of 82, we would report his performance as between 77 and 88.

If one wished to report test performance in terms of percentile bands, the percentile point corresponding to a particular raw score can usually be found in a table in the test manual. For example, a table might show that for a given set of norms a raw score of 88 corresponds to the 97th percentile and a raw score of 77 to the 81st percentile; this test performance would be represented by the percentile band for Reading Comprehension presented in Figure 2.

Using bands in place of single scores tends to decrease the possibility of giving too much importance to small differences that may be due solely to chance. Thus, only when there is no overlap between bands is it customary to conclude that the difference between the raw scores represents a real difference in ability. When bands overlap, we usually conclude that the difference between the scores is probably just due to chance. Such comparisons can be made between an individual's performance on two tests, or between the performance of two individuals on the same test.

An example is presented in Figure 2. From this figure we can conclude that the student is better in reading comprehension than grammar. We should not say, however, that she is better in grammar than in spelling, because the bands overlap.

It should be noted that although bands provide a simple way of making decisions about differences in test scores, we cannot attach probability statements to the conclusions drawn. (See Gulliksen, 1950, p. 20; and Doppelt, 1956).

Use standard scores, such as T-scores and stanines, whenever possible. Standard scores are useful if one wishes to make comparisons across tests that have been standardized on the same group. They can also be converted into percentiles for ease in interpretation, and to some extent avoid pitfalls common to other types of scores.

Many people, however, feel more comfortable with *grade equivalent (GE) scores*. This comfort, unfortunately, is based more on familiarity than measurement considerations. The interpretation of GE scores is not as simple as it appears; in fact, such scores can be misleading.

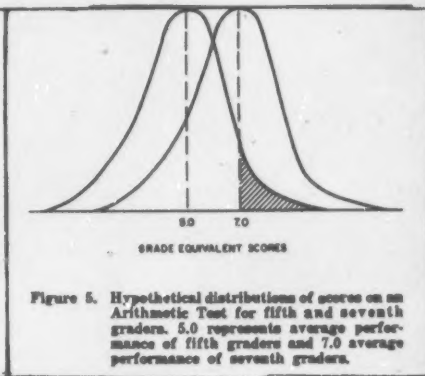
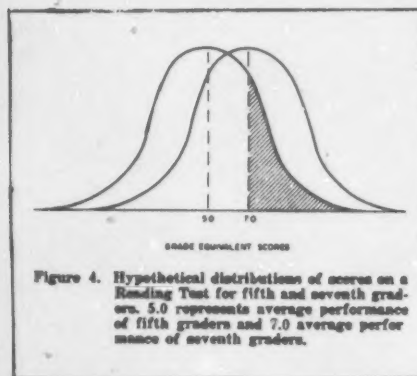
For example, when a fifth grader obtains a GE score of 7.0 on an achievement test in arithmetic, this does not necessarily mean that the student can solve the same mathematical problems a typical seventh grader can: the same questions are not usually given to fifth and seventh graders. Also, the GE scale is constructed on the assumption that learning progresses at a regular, steady rate. Not only is this assumption false, but the rate of learning is not always the same for different subjects (Beggs

and Hieronymus, 1968). Since GE scales are derived, to a large extent, through interpolation and extrapolation, GE scores do not necessarily represent the typical performance of students in a given grade. This is especially true at the extreme ends of the GE scale.

GE scores can also lead to confusion when one is comparing scores on tests measuring different abilities. A fifth grader achieving a GE score of 7.0 on a test of reading, as well as on a test of arithmetic, may only be the 80th percentile for fifth graders in reading, but at the 90th percentile in arithmetic. Why is this?

One difficulty is that the spread of scores differs from subject to subject, and from grade to grade. Also, there are differences in the extent to which distributions of scores overlap.

An example may serve here. The distributions in Figure 4 and 5 represent the situation where a GE score of 7.0 for a fifth grader only corresponds to the 80th percentile in reading but to the 90th percentile in arithmetic. Learning to read is to a large extent influenced by factors other than classroom experiences; learning arithmetic, however, is more dependent upon classroom teaching. Thus, one would expect the scores on the arithmetic test to be more similar within a grade than the scores on the reading test (see Figures 4 & 5). This is demonstrated in Figure 4 by the greater spread of scores on the reading test. It should also be noted that the score distributions overlap more for reading than for arithmetic, again reflecting the greater dependence of arithmetic-learning on classroom teaching.



In general one should avoid using GE scores. This is particularly true in high school where the rate of growth in many subjects tends to decrease considerably. There is, for example, a much smaller difference between the average reading ability of tenth and twelfth graders than there is between the average reading ability of second and fourth graders. The difference of two years on the GE scale would actually represent a smaller difference in reading ability for the high school students than for the younger students. In addition, GE scores have no meaning in courses that are taught in one semester or one year, e.g., biology and geometry. Another problem in interpreting GE scores for high school students is the different curricula followed by the students. For a more detailed discussion of problems involving GE scores see Angoff's chapter in Educational Measurement (in Thorndike, 1971, pp. 523-525).

Percentiles can be used effectively when reporting test results. Most people can readily grasp what is meant by the 85th percentile, if it is explained that the student surpasses 85 percent of those in a specified reference group. There is a problem, however, in that distances between percentile points can not be assumed to represent equal amounts of ability. Differences between scores at the extreme ends of a percentile scale represent greater differences in ability than do those near the mean. That is, there is a greater difference in ability between two students scoring at the 90th and 95th percentiles respectively, than between two students scoring at the 50th and 55th percentiles. The problem can be overcome by using standard scores, or by presenting percentiles in graphic form; the greater differences at the ends of the scale can be represented by leaving larger spaces between the percentile points (refer to Figure 2).

Since percentiles and GE scores are on an ordinal scale, they are not suitable for use in studies using statistical techniques requiring data on at least an interval scale. For this reason, scores are frequently converted to standard scores, such as T scores or stanines, where the distances between scores are assumed to be equal. Persons planning to do statistical analysis should consult statistics books on this subject.

Encourage realistic expectations. Teachers use standardized test results for two major purposes: to evaluate their students' performance in relation to the performance of other students, and to evaluate the curriculum.

Some teachers set as a goal that none of their students should score "below the norm." They are then quite unhappy when this goal is not reached. Such goals, by and large, are unrealistic and come from a lack of understanding of just what constitutes a norm.

The norm as used in this context refers to the average score of a group that has previously taken the test. By definition, 50% of the students in the norm group score below average and 50% above average. If the teacher's class is similar to the norm group then he should not be disappointed if

about half of his class does score below the norm. If all teachers in the country were successful in getting their students to score at or above the norm, the test publisher would probably renorm the test to give the scores meaning. The teacher would then be back where he started from.

Grade equivalent scores can create particular problems here. This is best illustrated by the following example from a published achievement test. A GE score of 5.6 corresponds to the average performance of students in the middle of fifth grade. This means that in the norm group 50% of the students in the middle of fifth grade received GE scores below 5.6. Therefore, the teachers in typical fifth grade classrooms should expect some of their students to receive GE scores at the fourth, third, and even second and first grade levels. Such low GE scores are an artifact created by the GE scale.

It is important, however, to know what norm group is being used for comparison. The norm group might be a national sample of students in a given grade, or perhaps a sample of students from the local school system who have taken the test in previous years. If the composition of a class differs markedly from the norm group on important characteristics (e.g., sex or grade or previous experience), then one would not necessarily expect the test performance to be the same for the two groups. Also, if the test content is not related to what is being taught in the school system, one might again expect differences. Small deviations from the norm should also be ignored, since they probably reflect only measurement error.

Perhaps one of the most beneficial uses of standardized achievement tests is in the identification of areas in the instructional program that need further attention. A useful approach involves the examination of test results for specific content areas or objectives on each subtest. In order to do this, the teachers need an item analysis by content areas or objectives. This can be provided as part of the output of a test scoring service on those standardized tests which have identified items according to specific objectives or concepts. Once the difficulty levels have been determined for each item within a cluster of items related to a specific objective or content area, it becomes easier to note differences in levels of performance between clusters of items. It is also necessary for teachers to decide whether each objective or concept is an integral part of their own instructional plan. Once this is done it becomes expedient to establish a performance level for each cluster group. One interesting approach is to ask the involved teachers, without benefit of test results, to establish a performance level for each concept and then reveal the actual results. If there are discrepancies between expected and actual outcomes, it then becomes the responsibility of the interest group to devise methods or programs designed to improve performance of their students in the noted areas. The limited coverage of content provided by the test must also be kept in mind.*

* Paragraph contributed by Anthony J. Allen

SUMMARY

In conclusion, the testing director should make every effort to see that the proposed uses of the test results are actually carried out. This means resurrecting test results from files, preparing written and oral reports, providing in-service training, and making practical suggestions based on the test findings and other relevant data. The testing director cannot wait to be asked, but must take the initiative in supplying information where and when it is needed.

REFERENCES

- Angoff, William; Scales, norms and equivalent scores, in R. L. Thorndike (ed.), Educational Measurement, 2nd Ed., Washington: American Council on Education, 1971.
- Beggs, D. L. and A. N. Hieronymus; Uniformity of growth in the basic skills throughout the school year and during the summer. Journal of Educational Measurement, 1968, 5(2), 91-97.
- Bloom, B. J., J. T. Hastings and George F. Madaus; Handbook on formative and summative evaluation of student learning. New York: McGraw-Hill, 1971.
- Doppelt, J. E.; How accurate is a test score? Test Service Bulletin No. 50, Psychological Corporation, 1956.
- Gulliksen, H., Theory of mental tests, New York: Wiley, 1950.
- Mathews, Walter M., Narrative format testing and traditional testing reports: A comparative study, Journal of Educational Measurement, 1973, 10(3), 171-178.
- Thorndike, R. L. and E. Hagen, Measurement and evaluation in psychology and education, 3rd Ed., New York: Wiley, 1969.
- Ward, A. W. and J. L. Mazur; Memo to a test director - No. 1: The test director in today's schools, NCE Measurement News, 1973, 16(2), 2-4.
- Wesman, A. G.; Double-entry expectancy tables, Test service bulletin of the Psychological Corporation, 1966, No. 56.

MEMO NO. 7: USE OF TESTS IN PROGRAM EVALUATION

by

James H. Lawson
University of Denver

and

Annie W. Ward
University of South Florida

In recent years there have been a plethora of "innovative," "experimental," "demonstration" programs as schools have tried to find ways to solve the many complex problems of society with which they are dealing. A great majority of these programs are funded through external funds, usually accompanied by a stipulation that the program be evaluated. Funding of continuation grants is usually based on the ostensible demonstration of effectiveness. The continuation of such programs from local resources after the funding source is withdrawn is also presumed to be based on the demonstrated value of the program. Furthermore, some school systems have begun to institute some form of data-based decision-making for local programs and activities.

The procedures and techniques of program and curricular evaluation are outside the scope of these memos. The reader who is concerned with these matters is referred to publications such as the AERA Monograph no. 1, Perspectives of Curriculum Evaluation (Tyler, et al, 1967); the April, 1970 issue of Review of Educational Research, and the many journal articles dealing with the subject.

This memo deals with the problems associated with the use of tests in evaluating special programs and curricular changes.

Most educators would agree that information is a better basis for decision-making than assumptions. The purpose of program evaluation is to provide information needed to make such decisions as whether to modify the new program, continue the new program or replace some other program with the new program. One part of the information which is needed concerns the effect of the program on students. Two types of questions may be asked:

1. *Does this program result in the improvement of student learning?*
or
2. *Does this program result in no loss of student learning, at a lower cost of money, personnel, time, etc.?*

Both of these questions require data as to the status of student learning after a period of program participation. Furthermore, both questions require information or assumptions as to what the status would have been without the program.

The decision as to the quality of performance of a defined group rests implicitly on a comparison with groups who have had competing experiences. In the absence of data about these groups, the evaluator is faced with making assumptions about what the situation might be. Surely some data, even though somewhat inaccurate, would be better than no data at all.

The preferred model for evaluating the outcome of a program is the "pre-test -- post-test experimental control group design," with samples randomly assigned to each group (see Campbell and Stanley, 1963). Inability to carry out this design, or failure to plan for the evaluation until the program has been operating for some time, frequently leads to an attempt to conduct the evaluation without either the control group or the pre-test.

Out of necessity, many ways of handling the evaluation under these circumstances have been solved. Sometimes it is possible to find an existing "comparison group" which is similar to the program group in many ways, and treat the comparison group as if it were a control group. When this is not feasible, many evaluators have turned to normative data on standardized tests as a substitute for control group data. Others have turned to "criterion-based" or "objective-based" assessment, in an effort to eliminate the need for comparison groups. All of these practices have certain pitfalls which are identified in this memo. Finally, there are certain problems associated with the use of both norm-referenced, standardized tests and locally developed, criterion-referenced tests which must be considered.

USE OF EXISTING COMPARISON GROUPS

If possible, students for whom a program is designed should be randomly assigned to treatments, so that one has a true experimental and control group. Since this is often not possible, the next best solution is often used -- the pairing of similar schools and random assignment of schools to treatments. This raises the question of the definition of "similar schools". Depending on the nature of the special program, these variables may be used to pair schools: mean and variance of IQ scores, mean and variance of achievement test scores, racial make-up and distribution of socio-economic levels. This solution is better than some others, but it still leaves areas of uncertainty, such as: are the variables used to pair schools the relevant ones, were the two groups of schools truly comparable at the beginning of the program, can one legitimately use data analyses which are based on assumptions of random selection? Since there are no definitive answers to these questions, the findings must be interpreted cautiously, and studies must be replicated.

Furthermore, small school systems may have serious problems in locating "similar schools". In both small and large systems it is sometimes possible to select a "quasi-control" group by sampling data files -- assuming that the system has collected a wide variety of data on its students on a systematic basis, and that some of the data are relevant for the project being evaluated. A group of students may be identified and used as blind controls, even though they attend a variety of schools.

USE OF NORMATIVE DATA ON STANDARDIZED TESTS

As indicated earlier, the normative data supplied by the publishers of standardized tests are frequently used as a substitute for control group data by evaluators. One of the first problems the evaluator encounters in using normative data is to determine whether his sample of students is comparable to the sample of students used in standardizing the test. If his student sample is not similar to the norm group as described by the test publisher, then the reliability and suitability of the test are questionable. The dilemma stems from the fact that normative samples are selected to give a wide distribution of scores, whereas most experimental programs are aimed at students whose scores are likely to yield restricted distributions. Furthermore, experimental groups are usually selected for programs on the basis of some characteristics which differentiate them from random representative samples.

Another sampling problem which the evaluator must resolve is the congruence or lack of congruence between the instructional objectives of his program and the objectives tested by a limited number of items selected by the test publisher. Using as an example a spelling test, most standardized tests can only devote fifty or so items to a spelling section. Assuming that there are five options for each item, it is likely that the standardized tests can present only 250 words. Whether these 250 words were included in the instruction is one question, and whether picking out misspelled words in a spelling test or a proofreading test fits the instructional objectives is still another question which the evaluator must answer.

Should the evaluator decide that the student and item samples match his student sample and instructional objectives, there is yet another obstacle the evaluator must overcome. Standardized tests are so designed that the average item difficulty is at or near the 50 percent level. This means that expectations of student performance on the average cannot be predicted to be better than 50 percent on a given item on the test. This of course, assumes the students taking the test are similar to those on which the test was standardized. Whether the evaluator wishes to accept the 50 percent level of performance for his group or not may depend on his realization that his group differs from the norming population in ways which are favorable for better than "average" performances.

If all of this were not enough to dissuade the evaluator from using normative data for his baseline data, he must also consider a rate of change, i.e., he must estimate the difference between the pre-test and post-test scores his students might have been expected to achieve from the beginning to the end of the program. Changes in the norms for the comparable time period are used to make this estimate. Since the norms group was selected to represent the total population, while the program participants are usually a subsample -- often the lowest 10 - 25%, it is not reasonable to expect "normal" gains for the group. What is it "reasonable" to expect? Obviously there is no clear-cut answer. One method which has been used is to compute a "percentage of normal" index at the beginning of the program and use that to compute a "normal gain" for a group. That is, if beginning 4th graders have an average grade score of 2.5, the index would be $2.5/4.0 = .635$. At the beginning of the next year, the expected grade score would be 63.5% of 5.0 or 3.18. This is an extremely questionable procedure, especially when the problems associated with grade scores are considered.

Although the use of standardized tests for the establishment of baseline data poses many problems, it does have some practical advantages. Most obvious is that the test is easily obtainable and the necessary booklets, answer sheets, and scoring keys are ready made. And, in spite of the voiced objections to standardized tests of any kind, there is a kind of grudging admiration that they are probably the best available index of pupil achievement.

However, it is better to simply report status and withhold any decision as to the effects of the program on that status. Data collection in subsequent years should be planned so that there is at least a comparison group, rather than trying to use normative data.

USE OF "CRITERIA" OR "OBJECTIVES" IN LIEU OF COMPARISON GROUPS

The difficulties in locating comparable groups have led some evaluators to develop ways of establishing baseline data and expectancies from the project goals and objectives and defining "success" as improvement over initial status. For example, if a behavioral objective were stated as follows: "90 percent of the students will spell correctly 100 percent of the words on a given list", the expected student achievement in such a statement becomes the baseline data for the evaluator against which he now compares his experimental data and determines the success of the program.

In most cases, this procedure requires the development of tests or other assessment instruments which are directly related to project objectives. Constructing this type of test assures a one-to-one relationship between instructional objectives and test items. However, the

Instructional objectives are based primarily on teachers' judgments about expected levels of performance, and it may be some time before the evaluator has enough data collected to determine the accuracy of such projections.

Obviously, the problem of generalizing test performance from samples which are not the same as those he is currently working with is not present except insofar as the initial judgment of the teachers may be based on their experiences with many and varied groups.

The use of performance-based tests for baseline data is often considered to circumvent the problems of difference scores and rates of change since the tests are designed to determine levels of performance and little attention is paid to the entrance or initial level of performance of the students. (See, for example, Cronbach, 1968.) Going back to our spelling test as an example, the performance-based test indicates that a given percentage of students spelled a certain percentage of the words correctly. However, the percentage of students' correctly-spelled words before the instruction is not known. Furthermore, we do not know what the level of performance would have been without participation in the program. So we do not know whether there has been an improvement as a result of the program. Obviously, it is necessary to empirically establish pre-treatment criteria as well as post-treatment criteria.

SPECIFIC PROBLEMS ASSOCIATED WITH THE USE OF STANDARDIZED TESTS IN PROGRAM EVALUATION

Lack of Congruence with Program Objectives

In general, standardized achievement tests are developed by a procedure which cause them to lag behind curricular developments. That is, they are developed by a procedure which insures that they reflect the most commonly delineated curricular goals and objectives *at the time* they are developed. These objectives may not coincide with the objectives of an innovative program or a recent change in a curriculum. Furthermore, standardized tests are designed to reflect general growth or attainment of objectives which are common to the majority of curricula, rather than attainment of a set of specific objectives of a given program. For some programs this is not a serious problem. For example, if one of the goals of the program is to "improve basic academic skills," then the standardized test may be a very appropriate instrument.

In choosing a standardized achievement test it is always important to assess carefully the extent to which the test is congruent with the educational objectives of the school system. The use of such tests in the evaluation of special programs also requires that the test be congruent with the goals of the program. In most cases standardized tests will need to be

supplemented with other instruments and with tests developed to measure the specific objectives of the program.

Inappropriate difficulty level:

Since many special programs are aimed at students at the lower end of the distribution of achievement, three phenomena of standardized test scales must be considered in conducting program evaluation:

- (a) the effect of chance (b) the "floor" and "ceiling" of the test and
- (c) the regression effect.

Most standardized tests use multiple choice items for which "chance" or the probability of "guessing" the correct answer is well established, i.e., $1/n$ where n = no. of options. It is quite easy to determine the "chance-score" for any given test by simply dividing the number of items by the number of options per item. For example, if a test has 40 4-choice items the "chance-score" = $40/4$ or 10.

When the chance score is computed for the subtests of most of the widely used achievement batteries, and when that score is referred to the norms tables, it is found that the normed score looks as if it has some meaning. In fact, the first two stanines are apt to be made up of students who score at or below chance. Since this is the target group of many programs, care should be used to avoid attributing any meaning other than chance to these scores. One should not rely too heavily on the reliability coefficient and the standard error of measurement reported by the test publisher if there is reason to believe the experimental group differs significantly from the sample used in standardizing the test. The primary reason for using a standardized test in this case is that it provides a ready pool of items which can be easily obtained and conveniently administered. Furthermore, the evaluator's use of the test data differs from the test publisher's. From the scores obtained using these items, the evaluator will establish his own baseline data for estimating difference scores; whereas, the publisher's purpose in standardizing the test data was to improve interpretation of an individual score.

The *floor* and *ceiling* of the test are other important considerations in using test data for program evaluation. Standardized tests are developed for several levels which overlap somewhat. No one level of the test can be used to measure all levels of achievement from beginning first grade through high school. There are usually two or three levels for primary grades, two for the intermediate grades and one for junior high level. Students who score near the bottom of the test may secure a "grade score" which looks as if it represents some real, if low, achievement. In actuality, this normed score is simply an "extrapolation" of the scale which has been developed and is not based on actual performance. Similarly, very high raw scores will be beyond the "real" points on the conversion scale. When the *floor* effect and the *chance* effect are both operating as they probably are in the testing of low achievers, no

real measurement is taking place.

One solution which has been proposed is that of "off-grade-level" testing, in which students are tested with a test level appropriate to their performance level, rather than the one taken by most students of their age and grade status. This procedure has much to commend it. Students may take a test which they can read and really respond to, so that raw scores reflect real achievement rather than just guessing. However, there are some problems associated with the practice. Norms established on the basis of on-level testing do not hold up when off-level testing is used; consequently, the distribution of scores for a school district may be lowered somewhat if off-level testing is widely used. However, for purposes of program evaluation, it is certainly preferable to use tests which measure some real aspects of achievement, even though such testing may make it inappropriate to use publisher's "standard scores" or other norms.

The *regression effect* is simply the tendency of the extremes of a distribution to regress toward the mean on a repetition of the testing. This effect, together with *chance* and *floor* (or ceiling) effects if normed scores are used, makes it easy to demonstrate "improvement" for an innovative program which is directed at a group who are low in achievement -- at least for the first year. This is a little more difficult with programs for "gifted" students and second-year programs.

Converted Scores

Since it is desirable to test both "before" and "after" a program, it is often considered desirable to use alternate forms of a test. Sometimes it is also desired to use different levels of the test. When either practice is followed, one is faced with the necessity of converting scores to some kind of common scale.

Grade scores are frequently used for this purpose because they appeal to the unsophisticated as being intrinsically meaningful. Unfortunately, the measurement problems associated with grade scores are unsolvable; e.g., unequal units, interpolations and extrapolations of the scale. (See Memo #6 in Chapter III). Some test publishers have developed "standard scores" to equate various forms and levels of their achievement tests, and this seems necessary. However, there is some doubt as to how adequately these scores work, particularly for the low end of the distribution.

If possible, it is better to stick to raw scores, using only one test or tests of equivalent items.

These suggestions may help avoid some of the problems discussed on the previous pages:

1. Select (or identify) a comparison group.
2. Try to avoid post hoc analyses. Better to simply assess status and withhold any decision as to the relationship of the status of program participants to the program.
3. Collect a wide variety of data on a wide scale on a systematic basis. If the data so collected are congruent with the purpose and goals of the school system, it may be possible to construct a "quasi-control" group by sampling from data files.
4. Divide the baseline data into groups such as high-middle-low achievers and compare high-middle-low achievers in the experimental group with the comparable baseline group. This helps the evaluator predict difference scores by comparing groups which are more likely to have similar distributions of difference scores. "Blocking" of this kind in other relevant variables might also enhance estimates of improvement of student performance. It is conceivable also that certain "blocks" of students might be receiving little or no help from the program and this would be important information for adjusting program procedures or goals.
5. In developing criterion-referenced tests follow these suggestions:
 - a. Prepare large item banks from which to sample.
 - b. Secure item performance data on a sample of non-program participants. Performance-based test items should be derived from as large a sampling as possible. Collecting data from groups other than those in the project would allow the evaluator to make inferences about test results with greater confidence, providing the groups are similar to the experimental group.
 - c. Calculate item difficulty and discrimination indices. This would allow for the construction of tests with known item difficulty. It has been suggested that pre-program and post-program groups be used to establish these indices. (Ebel, 1973).

REFERENCES

- Campbell, D.T. and Stanley, J.C. Experimental and Quasi-Experimental Designs for Research. Chicago: Rand McNally, 1963 (Reprinted from Handbook of Research on Teaching).
- Cronbach, L.J. Course Improvement through Evaluation, Teacher College Record 64:672-83, 1963.
- Ebel, R.L. Evaluation and Educational Objectives, Journal of Educational Measurement, 1973, 10, No. 4:273-279, 1973. Educational Evaluation, Review of Educational Research, Vol. 40, No. 2, AERA, April, 1970.
- Tyler, R.W., Gagné, R.M., and Scriven, M., Perspectives of Curriculum Evaluation. Chicago: Rand McNally, 1967.

MEMO NO. 8: THE USE OF STANDARDIZED TESTS TO EVALUATE TEACHING

Annie W. Ward

It is not at all uncommon in these days of "accountability" for administrators and the public to demand that scores on standardized tests be used as a basis to assess teachers' effectiveness. The Test Director then finds himself in a very difficult spot. If he argues against such a practice, his arguments may be interpreted as an admission that the standardized tests have no value and do not really reflect student learning. On the other hand, if he lamely provides the class-by-class summaries which are requested, he knows he is cooperating in a gross misuse of test results. What can he do? He can try, first, to do some educational missionary work on the issues and problems; and, if that fails, he can provide as much in the way of statistical adjustment as his data permit. He may also suggest some alternative ways of assessing teaching effectiveness.

ISSUES AND PROBLEMS

The complexity of teaching and assessment of learning. Teaching is a very complex procedure which takes place in a complex setting. The evaluation of teaching is very closely related to the evaluation of programs. In each case, one must identify and assess the effect of many concomitant variables. Teachers should be evaluated only for those things which they can control. Many variables affect what students learn and interact with the teaching process; these variables cannot be ignored.

Specific problems related to the use of standardized tests. Scores made on a standardized achievement test are sometimes recommended as the ultimate criterion by which to judge how well students have been taught. At first glance, this seems to be a commendable procedure. The scores are objective and it is easy to compare averages and ranges for various classes. What is often overlooked, however, are these points:

1. Standardized achievement tests reflect accumulated learning, not just what has been learned during a single school year.
2. Standardized achievement tests are designed to measure attainment of a general set of broad objectives, which may have only a slight relationship to the goals of instruction for a given class in a given year.
3. Students from differing cultural backgrounds learn differently and respond differently to various styles of teaching.
4. Evaluation of teaching must consider not only achievement gains, but also some of the "side effects." Looking only at achievement test scores may lead to ignoring certain effects which may condition future learning. Evaluation of teaching must consider, in addition to gains on achievement tests, such variables as students' liking for school and learning, their aspiration level, and their emotional adjustment.

5. Class expectations and interactions are closely related to achievement results for the entire class; that is, the "class ambience" may affect the learning of most members of the class. It is very difficult to demonstrate that one teacher's students have learned better than another's when proper procedures are used to control for other sources of variance, e.g., learning ability of students, motivation or aspiration level of students, amount of previous learning, amount of time allotted to instruction in the school day, etc. and when noncognitive effects are considered. To further complicate matters, when test scores are used to evaluate teachers, there is a high probability that teachers may invalidate the entire procedure by "teaching to the test".

STATISTICAL ADJUSTMENT OF DATA

Because so many variables other than teaching have been found to correlate highly with student achievement, some school systems have developed complicated "correction" procedures, which involve computing some kind of Index of expectation and then comparing the actual achievement with the expectation. The variables which are used most frequently in this process include: Intelligence test scores, a measure of Socio-economic status, and "pre-test" scores (such as last year's achievement scores.)

When these corrections are made for data obtained under non-experimental conditions, measurement errors and regression effects may mask or distort the true effects. As in any situation involving tests and test scores, results must be interpreted carefully with full allowance made for the potential sources of error.

ALTERNATIVE METHODS

It is obvious to most teachers and to those researchers who have tried to study teacher effectiveness that there is at present no simple, infallible, way to assess teacher competence. Certainly the comparison of teachers solely on the basis of their students' performance on standardized tests cannot be condoned. Alternative procedures are necessary.

The procedures should fit the purpose for the teacher evaluation. Sometimes the purpose is simply to meet a legal or policy requirement to "evaluate teachers annually". Often the expressed purpose is to "improve instruction." Occasionally, the purpose is to "weed out incompetents." Many problems evolve from failure to clarify the purposes or from attempts to use a single process for many purposes.

If there are no expected consequences, as in the case of evaluation to meet an empty requirement, the process used is of little concern. However, in those cases where some consequence is expected, the method used in the evaluation may be crucial to the attainment of the desired consequence.

If the desired consequence is improved student learning, presumably obtained through better teacher performance, then the evaluation process must be one that builds trust and encourages teachers to examine their own performance critically with no threat of official censure or dismissal. Test results can be used as a starting point in this process; that is, they serve to monitor the teacher's class and help to flag possible areas of difficulty. Teachers can be encouraged to try different approaches and methods and see how they work, with the test scores serving as a gauge. Changes in the student population of the classes or in other variables can be considered as well. In this way, test results can be instrumental in "improving instruction." The administrator can also offer observation and collegial team consultation for teachers who are working toward self-improvement.

The situation in which a teacher is on probation and/or for whom a decision must be made as to whether to retain or dismiss requires a very different course of action. Indeed, the procedures and type of evidence which are required in order to dismiss a tenured teacher make this an almost prohibitive activity. It is somewhat easier to evaluate beginning teachers before they attain tenure. In either case, it is often better to try methods aimed at improving performance.

An attempt to base a case for dismissal solely on achievement test scores of students would and should fail. Often, the reason for dismissal has little to do with student achievement; the reason may be behavior such as failure to take roll, meet classes on time, or take bus or lunchroom duty. If these actions are to be the basis for dismissal, they should be documented directly by careful observation and recording. Test scores alone are not likely to make a convincing case for incompetency and a really desperate teacher can manipulate test scores so that they provide erroneous information.

The Test Director should, therefore, insist that test scores not be used as the sole basis for dismissal of teachers. A great deal of discussion and education may be necessary in some cases and administrators may need some help in documenting their case in some other way. If test scores are used, then all possible statistical corrections must be applied in order to correct as many of the biases as possible.

CHAPTER IV

TEACHER-MADE TESTS: THE TEST DIRECTOR'S ROLE

Margaret E. Backman, Ph.D.

Director of Vocational and Social Science Research
ICD Rehabilitation and Research Center

The test director is concerned with evaluation in the school system. This refers not only to standardized group testing, but teacher-made tests, as well. In fact, the teacher-made test is one of the most important assessment devices used in evaluating students. For one thing, teacher-made tests effect students' lives considerably, since the results of these tests directly effect the student's grades. Thus, it is important that tests developed by teachers be developed properly.

Some prospective teachers in educational psychology and educational measurement classes are exposed to test development and the necessary psychometric properties of a good test. However, the majority of teachers have not had even a single course in testing. Even when teachers have had such a course, they often find that they do not have sufficient time to write and rewrite questions and do all the calculations and recording of data required to develop good tests. This is where the test director can be of great assistance. The test director can make a very direct offer to the classroom teachers to help with the writing of their tests. This offer should not be presented in such a way that the teachers think that they have not been writing good tests, or that the tests that they are giving will be scrutinized and criticized. The offer should be one of assistance.

One of the best ways to approach teachers is through in-service training programs or group meetings where the services that can be offered by the test director can be explained and discussed. Now what kind of help can the test director provide?

In in-service education, the following areas can be discussed at length: (1) the necessity for developing an item file, (2) the pros and cons of multiple choice tests as compared to the pros and cons of essay tests, (3) the proper length of tests, (4) the proper ways to write multiple choice items, essay questions, open-ended questions and matching items, (5) the best ways to develop scoring keys, (6) the characteristics and values of norm-referenced and criterion-referenced tests, and (7) the descriptive statistics used to analyze test results. These specific topics will not be dealt with here because they have been quite adequately discussed in a number of books, such as Appraisal Procedures in Secondary Schools by Donald J. Brown (1970), The Measurement of Classroom Learning

by Norman E. Gronlund (1968), and Mastery Learning: Theory and Practice, edited by James H. Bock (1971). And one mustn't forget the series of booklets in the Evaluation and Advisory Series published by Educational Testing Service, particularly the ones entitled Making the Classroom Test: A Guide for Teachers, and Short-cut Statistics for Teacher-Made Tests (1964). These are all very good resource books for developing the kinds of information the test director might wish to present in an in-service education program.

Another possible resource is the audio-filmstrip developed by the ETS that covers six steps in the development of tests: Why am I testing?; What should I test?; Whom am I testing?; What kinds of questions should I use?; How long should my test be?; and How difficult should my test be? Information on this filmstrip and supplementary material is available from ETS in Princeton, New Jersey.

As part of the in-service training, the importance of developing a central test file should be stressed after test files have been developed. It is much easier for teachers to build tests of specific difficulty levels, stressing specific content. It also simplifies the development of make-up tests. However, if a teacher thinks he is going to have to develop a test file alone, it probably won't be done.

The Test Director can help by being the resource person who maintains the file. This can be done by asking teachers to submit a copy of their tests to the test director. The test items should not be put directly into the file, because most questions, particularly multiple-choice questions, will need to be rewritten and edited. The test director can offer to edit the items in a new test and can help to see that the instructions are clearly written, that the scoring key is efficiently laid out, and that the answers appear on either the right or left-hand side of the answer sheet or on a separate answer sheet for easy scoring. The questions should not be crowded on the sheet; sometimes this happens because school systems, in response to budgetary considerations, limit the amount of paper teachers can have.

An ideal situation is for the school system to have access to an optical scanner for machine scoring of tests. The optical scanner saves time for the teacher, reduces clerical error in scoring, and also facilitates item analysis and computation of group statistics.

If the scanner has a mini-computer attachment or if computer services are available, many additional services may be offered to teachers; such as listing students scores either alphabetically or by student ID number, ranking the scores from highest to lowest, and printing a histogram showing the distribution of scores. Computer programs can give the difficulty of each item and the discrimination

index, which tells whether or not a particular item differentiates between those who know the material and those who do not; the programs can provide means, standard deviations, and standard errors of measurement. The computer print-outs can also be used to facilitate and maintain the card file; by clipping item analysis data from the computer print-out and pasting it onto the appropriate card in the item file, item information can be up-dated as an item is used over and over.

The Test Director should not be put off by the idea of using a computer if he is relatively inexperienced in this area. There are technical personnel and "canned" program packages available to help set up systems and programs. It takes some work in the beginning, but once the system is operational, it is quite worthwhile and not difficult to use. For information on the various kinds of computer facilities and optical scanning machines that might be used, the test director may wish to consult the second edition of Educational Measurement, specifically Chapter 8 (Thorndike, 1971). (See also Memo No. 5, Chapter II of this monograph.)

If a computer is not available, the Test Director can suggest another way of handling the labor situation involved in test analysis--using the students in the classroom. A very simple procedure, is described in Short-cut Statistics for Teacher-Made Tests, (ETS, pp. 6-10).

Another topic that should be discussed with teachers is whether their tests are to be designed to discriminate among students or to test mastery of subject matter. This usually creates very interesting discussions. But whether or not one can test for mastery depends to a large extent upon the philosophy and policies of the school, and whether or not teachers are permitted to give large numbers of A's and B's. Related to this, the test director could discuss with teachers the uses of summative and formative evaluation. (Mastery Learning: Theory and Practice, Block, 1971).

A word at this point about the stating of goals or objectives. Over the past years much has been said about objectives. That one should start out with objectives when one is designing a course or a test, seems important and rational; however, there is often resistance from teachers to spending time writing objectives. Frequently teachers get so bogged down trying to develop the objectives that they don't get any further. In getting tests written, one approach which has worked well is to encourage teachers to develop questions that they feel are important, and then look over the questions that they have developed to be sure that they have sampled all important content, that is, to analyze the test as to objectives covered. Additional questions can then be written for areas missed.

A very simple blueprint can be suggested which will have perhaps four to six cells. On the vertical axis the teacher could jot down various topics and areas covered in the course. The horizontal axis would indicate the "behavior" the student is to develop in relation to the topics. Several schemes for classifying these behaviors have been developed (See Bloom, et. al; 1956, Gagne, 1970). Most teachers find a complex classification scheme very difficult to use and like a simple scheme such as that proposed in the ETS filmstrip mentioned above: "Knowing", "Understanding", "Thinking". Another system might be "Simple Knowledge," "Simple Applications", "Complex Applications". A third system is "Facts," "Concepts," "Processes". There is no magic about the classification system used. What is important that teachers recognize the importance of "higher level objectives," i.e., those beyond simple recognition and recall, and that the scheme be simple enough to be facilitative without requiring a disproportionate amount of time. The total number of items wanted in the test would be decided upon; then the number of items representing each area and behavior material could be entered. Areas to receive more weight would be assigned extra questions. This is preferable to assigning more than one point to an important item. Whether it is prepared before or after test questions are written, a test blueprint can help teachers check the test coverage against the instructional objectives.

This memo has not attempted to duplicate what has already appeared in many books, but it is intended to guide the test director to some reference materials and procedures that may be useful for analyzing and writing classroom tests, and to suggest possible areas where teachers may need help.

One cannot overemphasize the importance of the teacher-made test. This is one of the problem areas that the test director must struggle with, so that the evaluation of students is done as accurately and reliably as possible.

REFERENCES

- Block, James H. Mastery Learning: Theory and Practice.
New York: Holt, Rinehart and Winston, Inc., 1971.
- Bloom, Benjamin S. Taxonomy of Educational Objectives, Handbook I:
The Cognitive Domain. New York: McKay, 1956
- Brown, Donald J. Appraisal Procedures in the Secondary Schools.
Englewood Cliffs: Prentice-Hall, 1970
- Educational Testing Service. Making the Classroom Test: A Guide for
Teachers. Princeton: Educational Testing Service, 1969.
- Educational Testing Service. Short-cut Statistics for Teacher-Made Tests.
Princeton: Educational Testing Service, 1964.
- Gagne, Robert M. "Educational Objectives and Human Performance."
In J. D. Krumboltz (ed.) Learning and the Educational Process.
Chicago: Rand McNally, 1965, pp. 1-24.
- Gronlund, Norman E. Constructing Achievement Tests. Englewood Cliffs:
Prentice-Hall, 1968
- Mehrens, William A. and Lehmann, I., Measurement and Evaluation in
Education and Psychology. New York: Holt, Rinehart and
Winston, 1973.
- Storey, Arthur G. The Measurement of Classroom Learning.
Chicago: Science Research Associates, 1970.
- Thorndike, Robert L. Educational Measurement. (2nd ed.)
New York: 1971.

CHAPTER V

COORDINATING EXTERNAL TESTING PROGRAMS

by*

Richard L. Poole

The National College of Education

One of the responsibilities of the Test Director is the administration and supervision of external testing programs. The purpose of this chapter is to examine this activity in order to clarify the specific tasks required at each stage of the administration process, and also to examine the ramifications for school policy.

DEFINING EXTERNAL TESTS

Several years ago Tyler (1963, p. 193-194) cited definitions of external tests from both the Joint Committee on Testing and the North Central Association's Committee on Articulation. The former group defined them as "those tests which are initiated by, or sponsored by, one or more agencies or organizations outside the local school district," while the latter group defined these tests as "those which are used primarily by some institution or organization other than the high school, and those over which the local school has, or feels it has, no real choice as to whether its students take these tests." Although by definition external tests are designed, initiated, sponsored and used by agencies outside of the schools, their target, for the most part, is the population of students in these schools. For most of these tests, students elect or choose to take them of their volition, although others are administered to selected students for purposes of establishing norms or baselines (e.g., national assessment, state assessment, norming of standardized tests).

Some of the reasons for the administration of external tests are: (1) to aid students in applying for training or educational programs external to those of the school, (2) to aid external agencies to admit, select, or place students in programs within their control, (3) to provide normative data for widely used tests, and (4) to provide data on the status of educational systems. Because these programs are directed at students in school, and because of the service orientation of the public schools, it is not unusual to have these tests administered by school personnel using school facilities. Schools and school systems need established policies and procedures for making such decisions as which external tests are to be administered

*with some comments and suggestions provided by Annie W. Ward, University of South Florida.

under school auspices, whether they are to be administered during the regular school day, and whether school personnel are to administer the tests as a part of their regular duties.

Guidelines for the approval and operation of external testing programs should be developed by the Advisory Committee. Within the framework of the guidelines, the Test Director of the school or school system should be authorized to make the decision as to participation in various programs. After the decision is made to include a given external testing program, consideration must be given as to whether it is to be considered a part of the regular school program or whether it is to be treated as something separate and apart. Finally, the Test Director must establish routines and procedures for handling the actual administration of the test. The remainder of this chapter deals with these points.

WHICH EXTERNAL TESTS SHOULD BE ADMINISTERED?

For some external tests, the local school or school system has little control over the decision of a student to take a test. However, it can control which tests are administered on school premises and/or by school personnel. Furthermore, schools may receive pressures of various degrees to encourage students to participate in a given program. The decision as to which external testing programs are to be sponsored by a given school or school system will be determined partly on the basis of how the administration of the tests will affect the school and its personnel, but the major consideration must be the effect on the students. Students should be protected from testing programs which are obviously exploitative; for example, those which have as their primary purpose the recruiting of customers for unaccredited or over-priced training programs. On the other hand, schools will wish to facilitate those testing programs which are potentially beneficial to students.

Questions such as the following should be asked: (1) When, where, to whom will the test be administered? (2) What is the test or testing program designed to do for students? (3) What is the test designed to do for the school or school system? (4) What is the test designed to do for education in general? (5) What are the costs to students and to the school in money and time? (6) What are the costs to students in loss of other kinds of activities?

Few external testing programs are designed to benefit students, the school, and external agencies equally. When the benefits of a program for students and the school are very obvious, there will be little hesitancy in deciding that the program is one in which the school should participate, even at considerable cost. However, most external programs will have limited benefits for the majority of students, and some of them will have no immediate benefits for either students or the local school or school system. It is for programs of this nature that general guidelines need to be established by the Advisory Committee.

REGULAR SCHOOL DAY, REGULAR DUTY or AFTER HOURS, EXTRA DUTY?

In resolving this question, the crucial issue seems to be the generality of the results. For norming of standardized tests and for state and national assessment programs, the benefits for individual students and for the school are minimal. Furthermore, there is usually no provision for paying either students or staff members for extra time. Therefore, if programs of this kind are to be administered, they must be handled within the framework of the regular school day. Fortunately, most of these programs have worked out procedures which minimize the amount of time which any one student or staff member must spend in the testing program.

In situations for which the benefits are primarily for individual students (i.e., selection testing for college admission or special awards programs) provision is usually made for compensating staff members who are involved in the test administration. If tests are administered during school hours, the compensation could be in violation of the contract which the individual staff member has with the school system, and the activity would be interpreted as a conflict of interest. Furthermore, the administration and supervision of external tests might intrude upon the instructional activities of the school. For these reasons, external tests which benefit primarily a few individual students should be scheduled after normal school hours or on the weekend. This is true whether a major external testing program is involved (such as SAT or ACT) or whether it is a special selection test which the guidance counselor is asked to administer for a single student as a part of the admission procedure for a private school to which the student is applying.

ADMINISTERING EXTERNAL TESTING PROGRAMS

Testing programs involving a small number of students are not too difficult to schedule and supervise, although care must be taken to maintain the security of the test materials and to see that the test is administered under standardized conditions. When the program is a very large one, involving many students, the test director is faced with a serious problem of logistics.

Some of the external testing programs are system-wide in nature and the system Test Director will handle all arrangements. Other programs are administered through individual schools with one person in each school taking responsibility for handling the program. This person should be clearly identified and assigned the authority to carry out all the external testing within the guidelines established by the Advisory Committee and under the general supervision of the Test Director.

Specific details of procedures will be very much dependent upon the size of the program, the number of schools within a given school system which are involved in the program, the policies of the testing agency, and the special situation at each school.

The rest of this chapter describes procedures which are appropriate for a large school and a large testing program which is school-based, rather than system-based. It is assumed that there is a "test supervisor" in each school. For testing programs which deviate from this situation, procedural changes would be needed, of course; however, this discussion will serve to point out the kinds of matters which must be attended to, and will suggest ways of handling them in the large-scale, school-based situation. It is hoped that these suggestions will indicate to the reader ways of operating which can be adapted to other situations.

ADMINISTRATIVE PROCEDURES

The process of test administration has at least three phases: (1) preparation, (2) testing, (3) follow-up. In the sections which follow, the characteristics of each phase will be identified in general, and then developed more specifically to make clear to the reader the types of tasks which each phase entails. In addition, the reader should realize that these tasks and phases are interrelated and that he will have to determine the best sequencing for his school system and not unthinkingly follow what is presented here.

Phase 1: Preparation

The activities of this phase start with some type of communication between the test publisher or sponsor and the test director pertaining to the feasibility of administering a particular test. And, assuming that it is permissible and possible to administer the test or tests in question, the activities go on to include those associated with the scheduling, allocating and managing of school facilities as well as the possibility of recruiting school personnel for test administration staff. The activities of this phase are complete when the test materials are carried to the examination room(s). The specific tasks of this phase cluster in the following four areas: (1) Scheduling Facilities, (2) Securing Personnel, (3) Conducting Public Relations, (4) Handling Test Materials.

(1) Scheduling Facilities

Since most school buildings are a bee-hive of activity, it is necessary to fit the testing schedule into the school calendar. This includes the scheduling of rooms, and articulating the schedule with other school functions. The type of event which such monitoring prevents is having a room scheduled for testing with an adjoining room scheduled for practicing cheerleaders. Other factors covered by such scheduling would be those of guaranteeing opportunities for ingress and egress, and alerting and informing building maintenance and security groups.

As a back-up measure, and for anticipating the unanticipated, particularly if the tests are scheduled for a Saturday or Sunday, the test supervisor should secure keys for the specific rooms and building scheduled to be used.

(2) Personnel

For large testing programs, the supervisor may need to arrange for other school personnel to assist with the test administration. Depending on the particular tasks these people are to perform they can be classified as a "room supervisor" or a "proctor". The room supervisor is responsible for all test materials during the time of testing and also for the reading of test directions to the candidates. Proctors assist the supervisor by performing such tasks as distributing and collecting test materials, replacing defective test material and monitoring the behavior of the candidates as they work through the test. Because supervisors have to spend time before the test familiarizing themselves with the contents of the test manual and have to account for test materials, as well as read the test directions and procedures to the candidates, they are paid at a higher level than the proctors. The types of tasks performed by both the room supervisor and the proctor together with their levels of compensation are determined by the testing agency or sponsor. In a large scale testing situation, it is recommended that one person be employed as a "stand-by" supervisor-proctor to be used in the event one of the scheduled individuals does not arrive.

(3) Public Relations

The process of testing sometimes raises public relation issues of both an internal and external nature for the school.

Internal issues are usually handled by using the normal procedures for scheduling and obtaining rooms. This permits the proper openness with regard to the articulation of the testing with other ongoing school activities. If the test is to be administered during regular school hours, teachers should be informed about the nature of the test, the purpose of it, and the procedures to be used. Even if the test is to be given after hours or during the weekend, teachers should be informed about the program so that they do not schedule conflicting events. An additional outcome associated with following such a procedure is that it may help to secure the support of teachers for the program and may provide an additional source of information for students.

External matters are more difficult to control. There is the possibility of miscommunication between the testing agency and the test director on the myriad of details associated with a test administration. In order to prevent difficulties of this kind, it is incumbent upon the test supervisor to double check all arrangements well in advance so that his school does not become the target for misplaced displeasure, anger or disappointment.

(4) Handling Test Material

If all the preliminaries have been attended to, then the incoming test materials should be promptly delivered to the test director's office without undue delay.

Upon the arrival of the materials, they are to be checked in against the accompanying packing slip or bill of lading. Materials should be checked for both their quantity and the identity of their specific serial numbers (test booklet numbers). Occasionally, the material will not check out and the director will have to telephone the testing agency to determine how he should proceed. After it has been ascertained that all the materials are in order, then they must be stored in a secure place until testing time. Some time before the actual time of testing, it may be necessary to reallocate, or pack them, for use by staff in separate rooms.

Testing in several small rooms at one time is a mixed blessing: On the one hand it reduces the volume of candidates and materials required per room and, possibly the consequences to those individuals if errors are made or material is lost. On the other hand, it increases the number of supervisors and proctors needed. For this reason, it is necessary to develop some sort of accountability system so that materials will not be misplaced or lost. Several systems are available, but to maintain the independent functioning of each testing room, each must be treated by itself. In those cases where a test roster of registrants is available, it may be used to assign candidates to rooms on an alphabetical basis. In other cases, it might be necessary to count out a subset of the entire group of candidates once they have arrived at a reception point or room. Regardless of what system is employed, it is necessary to inform the room supervisor as to who his proctors are and what test material he has. A simple check list should be prepared and used.

Also, if the test booklets are unnumbered, the test director should number them before they are reallocated for use in individual rooms. The reason for this is that such numbering and prepacking expedites counting. However, since the test booklets are not the property of the test director, he will have to number those booklets in an inconspicuous spot. Two such spots are the upper and lower corners on the folded or bound edge of the back cover of the test booklet. The point in using the folded or bound edge is, once again, ease of counting. Some readers may question the importance of such attention to the mechanics of counting and numbering of materials. However, the inconvenience associated with numbering test booklets prior to their actual use by candidates cannot be as impractical, in the long run, as the events which surround a lost or missing test booklet.

Additional items which may be packed with the test materials should be listed on the Sample Packing Slip. Of particular importance are the test manual, and the room floor or seating plan. Each room supervisor should receive the test manual some time before the test so that he can familiarize himself with the contents, the required test procedures and the explicit directions to be read to the candidates. This task can take several hours, and should not be left to the actual time of testing the candidates. Some testing agencies believe that their tests are unique, and even extend their uniqueness to the test manual by stating, "this test manual does not have a table of contents because we want our test supervisors to be familiar with all of its contents." Test manuals should be distributed to room supervisors at least one day prior to the examination so they may indeed read it. Although the time required by the room supervisor to read and become familiar with the test manual varies, it is something which cannot be done while attending to other testing activities or while teaching.

The other aspect of planning which is important in expediting the testing phase in each individual room is the floor or seating plan. Such a plan or chart helps the supervisor in formulating a testing strategy and in communicating it to his proctor.

Phase 2: Testing

Immediately following the end of the preparation phase the testing phase begins. The testing phase begins with the arrival of the testing personnel at the testing room, extends through the period in which the examinees are writing out or otherwise responding to the test questions, and ends with the departure of the testing personnel from the testing room. More exactly, the tasks of this phase focus upon three areas, namely: (1) Examinee admission, (2) Examinee-test interaction, and (3) Examinee dismissal.

(1) Examinee Admission Logistics

It is obvious that before candidates can be admitted to the testing room, they will first need to enter the building and then find their way to the specific room. In some large and complex buildings composed of many wings, this is no small feat. Hence, it is advisable for the test director to anticipate this and make several provisions for it. First, he may ask the testing agency to print on the ticket of admission to the test which doorway or building wing to enter. Second, he can have signs made up and posted to direct the students once they get inside the building. Third, he can inform his room supervisors of the number of candidates to expect, so if testing time is imminent and only a few candidates are in the room, the proctors may be sent to look for the missing candidates.

A rather formal procedure for admitting and seating candidates is recommended, in order to minimize the effects of any cooperative enterprises formulated by students to maximize their performance. As the students are checked in, they should be redirected to another proctor who has the task of directing students to particular rows or columns of chairs. This proctor should also alternately, or in small sequences of one or two students, direct the candidates to particular chairs. In addition, care should be taken to completely fill in each row or column of chairs in order to expedite the distribution and collection of test materials. When chairs are close together "completely filled in" means every chair to be used in accordance with the seating arrangement determined for a test in a particular room. It does not necessarily mean every chair in the room. After all the candidates have entered, each proctor should determine a head count for the candidates in the section assigned to him, and the supervisor should determine a head count for the entire room.

While the proctors are admitting and seating the candidates, the supervisor should be counting out the test materials for each section of the room in accordance with the number of students to be assigned to each row or column of chairs. If some chairs expected to be filled by candidates are left empty, then all the test booklets will not be distributed, and those that are distributed constitute a broken sequence of booklet numbers. The importance of completely filling all designated seats in each section is obvious in order to maximize the security and efficiency of material distribution and collection. A systematic method of handling materials to examinees is also very important.

It should be clear that as the proctors are handling the logistics of seating the candidates, they are preparing for the next task in this phase, that is, attending to the test booklets in such a manner as to keep themselves in control of the test booklets at all times.

After all the candidates are in their seats and before they begin to interact with the test materials, it is sometimes helpful to make some general announcements pertaining to the general testing situation. The range of such announcements would vary with the explicit test but could include items such as the following:

1. Introducing the testing staff.
2. Announcing what test is about to be given, and what candidates belong in the room.
3. Specifying the physical condition of the room with regard to heat and cold and how this might relate to candidates' dress.
4. Giving the location of rest rooms (and procedures to follow in leaving the room).

The end of these general announcements brings the examinee to the point of interacting with the test materials.

(2) Examinee-Test Interaction

The tasks associated with this aspect of the testing phase are by and large specified in the testing manual. By reading verbatim the directions found in the manual, the supervisor informs the candidates as to how they are to fill out their answer sheets, and how the test will proceed relative to such factors as starting and ending time, and the length of time for each test segment. The testing manual also specifies the order in which materials are to be distributed and collected, but not necessarily the procedures to follow to do this. Although the supervisor reads the directions verbatim, he may read them in such a manner that they do not communicate. The supervisor must learn to read directions so that they do, in fact, communicate. This requires that the supervisor read the directions slowly and that he pause to see that what was said is being implemented by the students. Occasionally, it may be necessary to direct a proctor to attend to an examinee who seems to need additional help. There will be tests or programs which have no manual, and in those situations the supervisor will have to formulate his own procedures.

One final item pertinent to this phase of testing is that associated with the disposition of individuals suspected of misconduct or cheating, as well as handicapped individuals. In regard to the former issue the supervisor should refer to the supervisor manual or seek assistance from the test director or the testing agency. Nevertheless, the supervisor should handle the matter with finesse and dispatch. In regard to the latter issue, handicapped individuals usually identify themselves to the testing agency and its representatives in turn make arrangements for an individual or otherwise special administration of the test. Special administrations vary with the particular test and testing agency, but may include events such as testing with a primary type edition of the test, or testing with a reader amanuensis. If a handicapped individual should present himself unannounced on the day of the test, then the supervisor will have to resolve the situation in terms of his available resources, and with minimum disruption or distraction to all candidates. In short, the handicapped individual will either be administered the test in the supervisor's assigned room with his leadership, or elsewhere with someone else supervising.

(3) Examinee Dismissal Logistics

Test materials are usually distributed by their order of use - pencils, erasers or pencil erasers, answer sheets, and the test booklet. It is easy to reverse this process for the collection process. Sometimes when this is done the attrition rate for pencils and similar objects may be quite high. The attrition rate is even higher if the candidates are instructed to hand in the pencils as they leave the testing room. If the high attrition rate is a problem, the test supervisor can employ at least two techniques to reduce it.

First, he can collect from the candidates all materials when they are still sitting. Secondly, he can base his collection procedure on the mind-set of the candidate. Examinees are usually eager to leave the premises as soon as the test is over. For the candidate the test is over when he hands in his answer sheet. For this reason, it is usually best to collect the answer sheets last. What is collected first depends on the trade-offs the test supervisor is willing to make. If he feels the test booklets are most important, or if he is fearful of losing one, he may first collect them. He does this assuming that the probability is quite small that a candidate will change any of his answers once the test questions are no longer available to him. He may, of course, decide that if the pencils are collected first, he'll get back most of them, and that the candidates will not have the means to change any answers even if the opportunity arose to do so. Either way the answer sheets are collected last, and the candidates appear only too happy to wait until they have been received and counted.

The final count of both the test booklets and the answer sheets are corroborated by the number of tests and answer sheets initially distributed. The number of test booklets and answer sheets initially distributed are corroborated by the head count of the candidates. Regardless of whatever procedure is adopted, it may be worthwhile to announce it in advance, and also that the candidates will be dismissed as soon as all materials are accounted for.

Following the collection of all the distributed test materials and the ascertainment of final and corroborating counts, there remains only the task of dismissing the candidates. Before this is done, it is wise to have the candidates look on the floor around their chair for any "misplaced" personal items, and to check their person for all personal items. The terminating task for the testing phase is the return of the test materials to the designated place.

Phase 3: Follow-up

The follow-up phase commences with the return of the test material to the collection center, continues with the separation of the test materials for mailing and return shipment to the testing agency, along with the completion of the paper work associated with the test, and concludes with the material being put into the mail, or into the hands of a carrier. One additional task which is sometimes performed is an analysis of the testing procedure to determine if anything needs to be changed.

The precise procedures for returning the test materials to the sponsoring agency are to be found in the supervisor's manual and need not be treated here.

Analysis of the test administration process can begin at any time following the test, and might start with an inspection of the Room Supervisor's Reports, or the observation of the test director as he monitored the entire activity. The point of such an analysis is not to impute blame, but rather to identify weaknesses and points of abrasiveness in order to synthesize a smooth testing process.

REFERENCES

- Tyler, R. W., The impact of external testing programs. In W. G. Findley (ed.), *The impact and improvement of school testing programs*. (The 62 Yearbook of the National Society for the Study of Education). Chicago, Ill: The University of Chicago Press, 1963, p. 193-210.

CHAPTER VI
REPORTING TEST RESULTS TO THE PUBLIC

by

James H. Lawson

University of Denver

The public learns of the results of tests administered in public schools through two modes. First, the results are reported to the individual child and/or the parents. Second, the status of student achievement of each school in an entire school district or state is frequently reported through the press to the residents of the school district, attendance area, or state.

The first mode of reporting test results might not be considered by some as public. However, the procedures employed by many school districts in this reporting process include supplying information about how an individual's test results compare with others in the same class, school, or district. In addition, an occasional parent or group of parents will object to the placement or categorization of their children and the results of testing and school district procedures become known to the public. Because the interpretation of test results to an individual can in various ways become public, it seems appropriate to discuss briefly such procedures.

The second mode of making test results known to the public deserves more attention since such reporting is by design intended to reach the entire community in order to appraise citizens of how well schools are doing in improving student achievement. Both school districts and state departments of education use these means in order to account to the taxpayers how effectively monies are being spent to support general or special education programs.

One cannot quarrel with the desire of school district or state department officials or the public to know how well students are doing in the schools, but one can find fault with the way the reports are made and what is reported. The general public is naive about how educational tests are constructed, and how tests results can be interpreted. Therefore, it is incumbent upon those responsible for reporting to the public to be especially scrupulous in their selection of data and the interpretation of those data to individuals and the various publics.

Reporting Individual Test Results

It is not the purpose of this section to serve as a substitute for a general text in tests and measurements. But it does seem appropriate to emphasize a few salient points regarding the interpretation of test scores and some general advice in reporting the interpretations. The reader should also refer to Chapter III of this monograph for additional suggestions on this subject.

The usual district-wide testing program is comprised of achievement and aptitude test batteries and the accumulated data attributable to each student are a series of test scores beginning with reading readiness tests in kindergarten, followed by achievement batteries through the primary and intermediate grades, and aptitude tests at junior and senior high school levels. Occasionally, interest inventories, general intelligence tests and vocational aptitude tests are also included in the testing program.

One of the most common methods of reporting individual test results is percentile rank because percentile ranks are fairly easy to interpret. Almost everyone seems to understand that percentile rank indicates the percentage of cases in a specified group who fall at or below a given score value. But what is not so easily understood or remembered is that one set of percentile ranks is rarely, if ever, comparable to any other. Furthermore, percentile rank is not an interval scale: this means that the percentile ranks for an individual on a test battery should not be averaged across the battery nor should the percentile ranks for a classroom, grade-level or any other group be averaged for a single test within a test battery. Therefore, each score value for an individual has its percentile rank and must be interpreted independently of other score values. In keeping with this same thought, if the test results for a transfer student from a district using a different test battery are recorded on school records, the percentile rank is not comparable to the percentile rank recorded for students taking a given district's test battery. Unless equivalent percentile ranks are available, percentile ranks on different tests are not comparable.

One major short-coming in using percentile ranks to interpret test results to laymen is that the concept of the standard error of measurement may be lost. However, two other forms of reporting test results may help convey the difference between an obtained score and a true score. They are test profiles and expectancy tables. Examples of test profiles are presented in Chapter III.

Expectancy tables equate score values to categories or different levels of a criterion variable. Frequently, high school level aptitude test results are used to predict college success. Expectancy tables help convey the idea of probability which accompanies the predictive efficiency of a score value as well as the possible range of scores associated with an obtained score.

In general it would seem that for reporting individual test results at the elementary school level the use of profiles containing percentile bands indicating the standard error of measurement might be the most accurate and least misleading method of reporting. The same would be true at the junior and senior high school levels for reporting individual achievement test results. For reporting aptitude test results at these levels the use of expectancy tables may be more appropriate.

Caveats

If there appears to be a discrepancy between an individual's achievement or aptitude test score and an intelligence test score, be slow to arrive at interpretations which lead to classifications such as "over" and "under" achievers. Remember each test has its standard error of measurement: no two tests correlate perfectly or nearly perfectly.

If there appears to be a discrepancy between an individual's preference or inventory test results and aptitude test result, do not be surprised. Correlation coefficients between these two types of scores have historically been exceedingly low.

If there appears to be a rather wide range in an individual's set of achievement test scores, do not be alarmed. Grouping students according to achievement scores might mean regrouping for every course or activity. Test batteries are designed to yield low correlation coefficients between sub-sets.

Other Achievement Reporting

Standardized test results are not the only kind of achievement results reported to the individual. The results of both teacher-made tests and course grades are reported to students and parents. Both of these constitute major reporting systems to the public.

Good teacher-made tests in many ways are more revealing as to the student's level of achievement than standardized tests. Teacher-made tests may be more closely related to the teaching objectives so that parents and students can quickly assess the level of mastery being achieved on a given examination. Furthermore, the comparison of a student's teacher-made test results with those of his classmates is more direct than comparing his scores with those of some ephemeral norm group.

Reporting School or District Test Results

The interpretation of an individual's test score on a norm-referenced test is tricky business at best. But there is some comfort in knowing the tests are primarily designed for the purpose of measuring

an individual's performance vis a vis a norm group. Furthermore, most publishers provide extensive information for the interpretation of an individual score. Unfortunately, the interpretation of group scores using most publisher's manuals is fraught with problems.

National Norms and Grade-Equivalents

It is a common and distressing practice of school districts and state departments to compare a school's standardized test results with those of other schools within the district and to compare the district's results with those of other districts, or with national norms. School-mean test results are generally expressed in grade-equivalent scores and commonly published in the following manner: "Sixth-grade students in our district scored almost a full year above the national average in language skills." Perhaps a better understanding of two measurement concepts might alleviate such malpractices. The concepts are "national norms" and "grade-equivalent scores."

First, consider "national norms." Each test publisher standardizes his test following prescribed procedures using a norm sample. This group of students is selected with the hope that it is representative of all students for whom the test is designed. For example, if the test is to be used with fourth-graders, the test publisher seeks a norm sample representative of all fourth-graders in the nation. Chances of obtaining 100 per cent representativeness are slim indeed. That is why most publishers describe the norm sample in great detail. The test-user may then determine whether he believes the test was normed on a sample which is representative; and even more importantly, whether the norm sample is similar to the students he wishes to test.

This latter point is very important, since it is axiomatic that a test should be used only with a group that is similar to the sample used to norm the test. That is, an individual's score must be compared to scores obtained from students who are similar to the individual. No one would insist that a fourth-grade student's score be interpreted on a scale derived from a test administered to students of some other grade level. Yet interpreting the score of an individual fourth-grader who lives in a large urban area with a norm table based on the scores of students living in some distant state is commonly insisted upon, in spite of the fact that the students from the distant state are probably no more representative of the students in the large urban area than are students of some other grade level. It is apparent then that norms are not standards nor truly national in nature; the "norm" for a standardized test is merely the average score made by pupils in the norm sample.

The second concept to consider is grade-equivalent scores (sometimes called "grade placement," or "grade norm" or simply

"grade scores") Grade-equivalent scores are not equal-interval scores. Since rates of educational growth and human development are not linear, grade equivalents are arbitrary scale values. This means that the amount of "reading achievement" represented by a gain from 4.1 to 4.5, for example, is almost certainly not equal to that represented in a gain from 6.5 to 6.9. In other words the grade equivalent scale might be compared to a yardstick on which some "inches" were 3 inches long while others were only 1/16 of an inch. Thus, measurements which use different sections of the measuring scale would not be comparable. For additional discussion of this problem, refer to Chapter III of this monograph.

Alternative Reporting Methods

Since using school-means and grade equivalents are not appropriate for reporting school or district student achievement, what alternatives are available? First, let's consider the reporting of criterion test results and second, matrix sampling procedures.

Criterion Test Results

Many school districts are now using some "criterion-referenced" or "objectives-referenced" test, either by choice or by mandate. This being the case, there exists a vast pool of data which for the most part is not being made public.

Assuming that the criterion-referenced tests are constructed for all grade levels within a district, comparison of levels of mastery within schools and across schools is possible. For example, if a fifth-grade social studies criterion-referenced test is constructed and administered to all fifth-grade students in a school district, it would be possible to report the per cent of students answering a given number of items correctly. Thus, one might report that 90% of all fifth-graders within the district answered correctly 75% of all the items on the social studies test. Such a procedure would necessitate either reporting or making available the test questions so that the public would have some idea about the face validity of the examination. An alternative might be making public sample questions in order to protect the confidentiality of the test if it is to be used again.

The decision to use objectives-referenced tests rather than normed referenced tests is often made under the mistaken idea that objectives-referenced test interpretation can be done with no need for a norms base. However, sober thought indicates that there are no a priori bases for setting expectations for student performance except that of a normative nature. That is, what a school or community expects fourth graders to know is based on their perceptions of what 4th graders do know. This point is frequently overlooked by advocates of "objective-referenced" testing.

Criterion-referenced testing obviously eliminates the comparison of student achievement between districts unless the school districts were to agree to the same learning criteria and use the same criterion-referenced tests. One criticism of this kind of reporting is that little is known about how well the students in one school or district are doing compared to "national norms", other school districts, or the students in last year's class. An answer to such criticism might be that unless large amounts of money are spent and much pre-planning is done to collect appropriate data, it is impossible to make such comparisons.

Another problem is that there is little agreement among experts about the psychometric properties of criterion-referenced tests, so the practitioner would do well to limit his reporting of test results to percentages, ranges, and median scores. Definitive statements cannot be made about the reliability, validity, or additivity of items of criterion-referenced tests. Also, it is apparent the reporting of criterion-referenced test results could be an overwhelming task if every subject and grade level were reported yearly; therefore, reporting of selected grade levels and subjects are usually scheduled for reporting over a period of several years. However, not all grade levels or subjects need to be included in any reporting schedule. The selection and scheduling would be an individual school district's decision and might be predicated on community concerns.

Matrix Sampling

If resources are available and some comparisons between schools and/or school districts is desired, matrix sampling procedures would allow school officials to make public the evaluation results of special or regular programs in a more precise manner.

Two excellent sources which describe the procedure in some detail are Principles and Procedures of Multiple Matrix Sampling by David M. Shoemaker and Chapter 8, "Introduction to Matrix Sampling for the Practitioner" by Kenneth A. Sirotnik in Evaluation in Education: Current Applications edited by W. James Popham.

Briefly, matrix sampling procedures allow the researcher to sample items as well as examinees, thereby eliminating the administration of all test items to all students. The resulting matrix (students and item responses) are amenable to statistical treatment such as estimating the standard error of the difference between means of scores of two groups.

The single most important characteristic of matrix sampling procedures is that it is geared to the assessment of group differences rather than individual differences and hence is suitable for comparing group performances.

The two other major advantages of this procedure are the reduction of testing time and the number of examinees needed for testing. Both of these are obviously important when collecting data from a single large district or several districts and when collecting data for many subject areas.

Although matrix sampling procedures may be somewhat complex for practitioners, the public can certainly understand reporting that states group "A"'s mean score differs from "B"'s at given confidence levels; and educators can feel better about the appropriateness of the procedures in making inferences.